中国精算师资格考试用书

精　算　模　型

Actuarial Models

主编　肖争艳

主审　孙佳美

中国财政经济出版社

图书在版编目（CIP）数据

精算模型／肖争艳主编．—北京：中国财政经济出版社，2010.11
中国精算师资格考试用书
ISBN 978－7－5095－2554－8

Ⅰ．①精…　Ⅱ．①肖…　Ⅲ．①精算学－资格考核－自学参考资料
Ⅳ．①F224.0

中国版本图书馆 CIP 数据核字（2010）第 200527 号

责任编辑：陈　冰　　　　责任校对：张　凡
封面设计：耕者设计　　　版式设计：兰　波

中国财政经济出版社出版

URL：http：//www.cfeph.cn
E-mail：cfeph@cfeph.cn

社址：北京市海淀区阜成路甲 28 号　邮政编码：100142
发行处电话：88190406　财经书店电话：64033436
北京富生印刷厂印刷　各地新华书店经销
787×1092 毫米　16 开　27 印张　652 000 字
2010 年 11 月第 1 版　2010 年 11 月北京第 1 次印刷
定价：58.00 元
ISBN 978－7－5095－2554－8／F·2173
（图书出现印装问题，本社负责调换）
本社质量投诉电话：010－88190744

中国精算师资格考试教材

编审委员会

总　　序

ZONGXU

精算起源于保险业，是保险公司经营不可或缺的核心技术之一。保险公司只有运用精算技术进行保险产品定价、准备金评估、风险管理等，才能在科学基础上实现保险业务的稳健经营，有效防范风险。

我们常说的精算，包括三个方面，即：精算理论技术、精算规则和精算师资格认证。

精算理论是对保险业务经营中各种不确定因素和风险规律的认识，精算技术以精算理论为指导，是精算工作中对各种不确定因素和风险进行识别、评估、定价、处置等所采用的方法、技术，包括所使用的数学模型、数学工具等。随着保险业经营实践的发展和人们认识的深化，精算理论技术也在不断发展。精算理论技术属于学术研究的范畴，可以存在不同的观点和流派，各种不同观点和流派之间的讨论、交流，可以促进精算理论技术的发展。

精算规则，是保险监管机关制定或认可的关于精算工作应当遵循、遵守或采用的原则、方法、标准、制度等规范。制定精算规则，以精算理论技术为基础，又要综合考虑一定时期的经济环境、保险业发展状况和风险特征、精算技术力量、监管政策的要求等多种因素。

精算工作需要专业人员从事，精算师就是具备精算的知识、技能，从事精算工作的专业技术人员。虽然精算师的从业范围不限于保险业，但主要还是在保险及相关行业就职（如对保险公司的精算报告进行审核的会计师事务所，为保险公司服务的精算咨询公司等）。在保险公司中，精算师责任重大。因此，必须经过资格认证，才能担任精算师（如同律师、注册会计师需要资格认证）。在国外，精算师资格的取得一般有两种方式：一种是通过专业资格考试取得，另一种是经过学历教育后取得，但主流是通过考试取得。在发达国家，精算师有自己的专业团体——精算师协会，一般由精算师协会组织资格考试，对通过考试的人授予精算师资格。

精算理论技术、精算规则、精算师资格认证三者相互联系，密不可分：精算理论技术是基础，制定精算规则、考试认证精算师，均以精算理论技术为基础，精算规则是精算师从事精算实务的直接依据。

我国自1980年恢复办理国内保险业务之后，曾长期缺乏精算专业人才，既没有制定精算规则，也没有建立自己的精算师资格考试认证制度。1988年南开大学在北美精算协会的支持下开办精算专业教育，此后国内又有多所大学开办精算专业教育，培养了一批精算人才。由于当时中国没有精算师资格认证制度，这些国内学习精算的人员主要是考取北美和英国等国外的精算师资格。1992年，国内的保险市场对外开放，外资保险公司进入国内市场，一些具有国外资格的精算师到国内工作。

1995年颁布并施行的《中华人民共和国保险法》中，要求寿险公司必须聘用经金融监管部门认可的精算专业人员，建立精算报告制度。《保险法》首先要求寿险公司聘用精算师、建立精算报告制度，是因为：第一，精算起源于寿险业务经营，精算技术在寿险业的应用较为成熟；第二，寿险业务期限长，风险更具隐蔽性，对精算技术的运用更为迫切和重要，第三，在精算专业人员严重不足、精算规则空白的条件下，同时要求寿险业和非寿险业聘用精算专业人员、建立精算报告制度，难以实现。

为此，当时的保险监管部门——中国人民银行保险司于1997年10月启动了“中国精算制度建设”研究项目，决定建立中国的精算师资格考试认证制度，并逐步制定精算规则。中国的精算师资格考试认证制度，主要借鉴北美精算协会的考试体系，把精算师资格分为准精算师和精算师两个阶段，分别设立考试课程，通过准精算师考试课程的，授予准精算师资格，在获得准精算师资格基础上，再通过精算师资格考试的课程，授予精算师资格。在课程设置、考试内容、难度等方面，均力求达到与发达国家的精算师考试相当的水平。在制度设计、拟定考试大纲、教材编写过程中，得到国际精算团体的大力支持和帮助。1998年11月，中国保监会成立之后，继续推进精算制度建设。2000年，中国精算师资格考试开考，与此配套的教材也陆续出版发行。中国保监会1999年发布了关于寿险公司的精算规定，建立了寿险公司精算规则体系的基本框架。

2002年10月《保险法》进行了第一次修改，于2003年1月1日起施行。修改后的《保险法》把聘用经金融监管部门认可的精算专业人员，建立精算报告制度的要求扩大到非寿险公司。因此，经过论证、筹备后，自2004年开始进行非寿险精算师的资格考试认证，称为中国精算师（非寿险方向），与此相适应，以前的精算师则称为中国精算师（寿险方向）。同时关于非寿险精算的规则也由中国保监会陆续制定发布。

2007年，中国精算师协会成立，组织精算师资格考试是协会的重要职能之一。协会设立了考试教育委员会，负责精算师资格考试和后续教育事宜（此前是由中国保险行业协会的精算工作委员会负责精算师资格考试）。

中国精算师资格考试施行10年来，通过考试认证了一批中国精算师和

中国准精算师，取得了一定成绩，积累了一定经验。目前已在北京、上海、天津、广州等 15 个城市设立了考试中心，并在香港、加拿大滑铁卢大学设立了 2 个海外考试中心，每年春秋两季举办考试。

随着国内保险市场的发育、精算技术的发展及国际精算界的变革，原有的考试体系已不完全适应。为此，中国精算师协会于 2009 年决定对中国精算师资格考试认证体系进行调整，并于 2011 年实施。调整的基本内容是：精算师资格考试仍分为准精算师和精算师两个阶段；在准精算师阶段，不再区分方向，对原寿险和非寿险两个方向的考试课程进行整合，考生通过 8 门必考的准精算师考试课程，并经过职业道德培训后，可获得中国准精算师资格；精算师则继续分为寿险和非寿险两个方向，有 3 年以上工作经历的准精算师，通过 5 门精算师考试课程，并经过职业道德培训后，可获得中国精算师（寿险方向）或中国精算师（非寿险方向）的资格，5 门精算师考试课程，既有必考的，也有选考的，具体科目，因寿险和非寿险方向有所不同。

对于在旧考试体系下已经通过的考试科目，如何转换为新考试体系的相应科目，也进行了研究，制定了转换规则。

为编写新考试体系的教材，中国精算师协会成立了教材编审委员会。教材编写力图贯彻国际性、先进性和实用性三个原则。国际性是指，鉴于中国精算师协会已正式申请加入国际精算师协会，因此精算师资格考试必须符合国际精算师协会的要求，达到国际精算师协会的标准。所以，在课程设置、课程内容、必考科目等方面，均以国际精算师协会的要求为标准。先进性是指，尽可能把精算理论技术的最新成果包括在这套教材之中。实用性是指，教材内容紧密联系国内保险业的实际，考虑国内精算人员需要掌握的知识和技能。

教材的具体编写实行主编负责制。教材编审委员会研究、协调、决定教材编写中的重大事项，确定各门课程的主编和主审人员，指定协调人对若干相关课程的内容调整、取舍和进度进行协调。教材初稿完成后，不仅由主审进行审阅，而且组织保险公司的相关人员进行试读，提出修改意见。教材的主编、主审、试读人员，都是在保险业、精算界具有业务专长、经验较为丰富、具有一定影响力的人员。可以说，这套教材的编写，是集中了行业的智慧和力量，凝结着组织协调人员、编审人员、试读人员的心血。

尽管如此，我们仍不认为这套教材已经尽善尽美。由于经验不足、认识水平有限，也由于时间仓促，教材在某些方面还显粗糙，还存在许多可改进、待完善之处。我们希望在教材投入使用之后，听取专家、考生和社会各界人士的意见，将来进一步修订。

回顾中国精算师资格考试 10 年来的历程，是在保险监管机关的领导

下，在保险业、有关高等院校及社会各界的积极参与下，在国际精算组织的支持下，不断发展、完善，取得进步的。在此，我谨代表中国精算师协会，对多年来关心、支持、参与、帮助中国精算事业发展的有关领导、专家和广大的精算专业人员表示真诚的敬意和感谢！

中国精算师协会　会长

2010 年 11 月 15 日

编写说明

BIAN XIE SHUO MING

精算科学的主要部分是构造和分析数学模型，这些模型刻画了保险赔付损失，以及资金流入和流出一个保险系统的过程。精算风险也可以用随机模型的方法进行表述，这些模型是对这些精算风险变量未来的概率分布及环境状况的假设。本教材的目的是向读者阐述精算建模的过程，即如何从实际数据出发建立一个合适的精算模型。

长期以来，我国精算课程和考试体系都包含了精算建模的内容，但是它们都分属于不同的课程体系。生存模型及估计是寿险精算的重要基础，用于确定身故、失效和伤残时间的概率模型，这部分内容被放在“生命表基础”课程中；理赔额和理赔次数一般模型是非寿险精算的重要基础，用来确定非寿险公司的赔付损失分布，这部分内容被放在“风险理论”课程中。虽然生存模型与理赔量和理赔数分布模型刻画的风险不同，但从统计方法上，生存模型的研究与理赔量和理赔数的一般模型并没有本质的差异。因此，国际精算师协会（IAA）在国际精算教育指南和国际认可精算师资格考试的培训大纲中，将这些内容放在一门课程（IAA6）中。为适应 IAA 考试体系的要求，中国精算师协会考试委员会将中国精算师资格考试课程“05 风险理论”和“06 生命表基础”整合成新课程体系中的“A3 精算模型”。作为这门课程的指定教材，本书试图将这些内容整合在一起，从精算建模的角度出发，以概率统计为研究工具，对保险经营中的损失风险和经营风险进行定量的刻画，建立精算模型并研究模型的性质。

为了保持与原考试课程体系的连贯性，本书是在中国精算师资格考试用书《风险理论》（吴岚、王燕主编，中国财政经济出版社 2006 版）和《生命表基础》（李晓林、孙佳美主编，中国财政经济出版社 2006 版）的基础上进行编写和修订。根据 IAA 课程大纲，本书增加了多状态生存模型、理赔额和理赔次数的分布、布朗运动与盈余过程、信度理论、Bootstrp 模拟和 MCMC 模拟等内容；保留了生存模型、生命表、短期个体风险模型、长期聚合风险模型、修匀理论、随机模拟等章节的基本内容，删除了人口统计和效用理论等内容。本书的最大特色是，重新编写了三章的内容来阐述在完整和非完整样本数据情况下生存函数、理赔额和理赔次数分布模型的估计和选择，并用两个案例来说明整个精算的建模过程。

虽然本书是精算师考试教材，但并不要求读者对保险系统已经具有很

好的知识背景。凡是在本书中首次出现保险术语的地方，我们都会给出定义。本书同时也是一本统计应用教材，既适合于具有中等概率统计知识的读者学习怎样运用统计学知识处理和研究保险业务的问题，也为已经掌握较多数理统计和随机过程知识的读者提供较深的理论内容，以便更好地掌握保险精算知识。

本书由中国人民大学统计学院肖争艳老师担任主编，负责全书统稿和编写第一、八、九、十、十二、十四章，并对第十一、十三章的初稿进行了修订；中央财经大学保险系郑苏晋老师负责编写第二至第四章，并对第五至第七章的初稿进行了修订；北京大学数学科学院的吴岚老师作为教材协调人，对全书的大纲和内容提出了方向性的指导意见；南开大学孙佳美老师作为教材主审，对全书进行了认真细致的审阅，并提出了许多宝贵的修改建议；李晓林、贾冬梅、钟颖、史淼、郭程宁等作为试读人，为本书进行了认真的评审，为保证全书的出版质量提供了有力的支持；北京大学数学科学院的杨静平老师也对本书提出了一些参考意见。在本书的编写过程中，中国人民大学统计学院的一些研究生也参与了本书初稿编写和习题解答等工作，他们是：邵亚娣、李君、刘天营、王伟伟、左辰、鲍金辉、张逸铭、徐梦语、蒋安华和程夏莹等同学，同时还有许多读者、专家也提出了宝贵的意见和建议，在此一并表示衷心的感谢。

编者

2010年8月

目 录

第一章　绪　　论

学习目标

□ 了解精算建模的一般过程

□ 了解参数模型与经验模型的优缺点

□ 了解本书的基本结构

§1.1　构建精算模型

所谓模型，就是对现实的一种数学简化。对任何给定问题的研究，都可以用模型化的方法来解决。精算中许多问题的解决都需要借助于模型。例如，在寿险中，通过建立生存模型，对人口的死亡规律进行分析来预测被保险人未来的赔付；在非寿险中，精算师通过估计被保险人的索赔次数和索赔额的分布来进行费率厘定、准备金计提、再保险安排等一系列精算问题。因此，北美精算协会在公开发表的《精算学基本原理》中指出："精算风险可以用随机模型的方法进行表述，这些模型是对这些精算风险变量未来的概率分布以及未来的环境状况的假设"①。这里的精算风险变量一般指：风险是否发生、发生的时间和损失量——索赔事件的发生机率、如果索赔事件发生其发生的时间以及围绕索赔的所有成本。

精算模型的构建有两种方法：经验法和参数模型法。经验法就是不对模型做任何分布假设，直接使用经验数据建模。当统计数据特别充足而且完整时，经验分布趋近于真实分布。但在通常情况下，我们所获得的样本数据是有限的，尤其是关于高额赔付的数据更为有限。有时得到的数据还是不完整的，有可能被截断或被删失。这些都会导致经验法存在偏差。下面两个例子将阐明这一点。

【例 1－1】 某团体人寿保险合同由不同年龄和不同受益水平的 500 个雇员组成。在过去的 5 年中，已有 8 名雇员身故并共计得到 45 万元。由于该计划的身故赔付与雇员的工资水平挂钩，所以需要将赔付进行通货膨胀调整。假设下一年通货膨胀率是 10%，试根据以上信息对该合同下一年的预期身故赔付进行经验估计。

---

① 这句话来自 Society of Actuaries Committee on Actuarial Principles, "principles of actuarial sciencc", Transactions of the Society of Actuary, 1992, 第 571 页的原理 3.1。

解：5 年内年平均赔付额为 9 万元，考虑到通货膨胀因素，预计下一年预期身故赔付为 9.9 万元。当然这个估计的缺陷在于，过去 5 年的经验不一定完全能够反映这个合同在未来一年的情况，因为在如此短（5～6 年）的时间内身故赔付的表现可能会有很大波动。 ■

看来，更合理的方法是建立一个模型。依例 1－1，应建立一个生命表，而要构造这样的表需要积累很多个体的经验，500 个人的经验是不够的。有了这张生命表，不仅可以估计下一年的预期赔付，还可以度量我们所作的估计本身的风险。

【例 1－2】 考虑一个公司团体牙医保险计划。目前保单规定，每次事故的免赔额为 50 元，即只对一次医疗费用超过 50 元的保单赔付超出的部分。为了减少保险公司的平均赔付成本，有 3 种修改方案。第一种方案认为应该取消免赔额，这样员工就会经常去看牙，从而减少高昂的医疗费用；第二种方案认为应该提高免赔额到 100 元，以降低赔付成本；第三种方案认为应该限制对高额损失的赔付，建议保持免赔额 50 元不变，但每次最高理赔额不超过 2 000 元。作为精算师，你认为哪种方案比较合理？为了研究方便，假设你已经随机抽取了 10 个赔付数据：141、16、46、40、351、259、317、1511、107、567。

解：在免赔额为 50 元的条件下，每次赔付的平均值为 335.5 元。如果免赔额提高到 100 元，上述保单的赔付额数据将变为：$141-50=91$，$351-50=301$，…，$567-50=517$ 元，其中赔付额低于 50 元的保单将为 0。于是平均理赔额为 $(91+301+\cdots+517)/7=2\,903/7=414.7$ 元。保险公司的成本将减少 $(3\,355-2\,903)/3\,355=13.47\%$。

当免赔额为 0，上述理赔额分别为：$141+50=191$，$16+50=66$，…，$567+50=617$ 元。平均理赔额为 $3\,855/10=385.5$ 元。保险公司的成本将提高 $(3\,855-3\,355)/3\,355=14.90\%$。

实际上，当取消免赔额时，经验法得到的结果是没有意义的，因为我们使用的数据并不是真正来自原始样本。如果免赔额为 0，则任何有医疗花费的保单都可以获得赔付，也都应该有可能被随机抽取到。但是由于最初的保单规定免赔额为 50 元，因此医疗费用低于 50 元的保单损失不可能被保险公司所记录，更不可能被随机抽取。这样会损失了大量的原始数据，造成估计结果的偏差。

对于第三种方案，经验法无法衡量这个修改对平均赔付额的影响，因为样本数据中没有理赔额超过 2 000 元的数据。

经验法的上述缺陷可以通过建立参数模型来解决。在例 1－2 中，假设每张保单的原始医疗费用服从对数正态分布 $LN(\mu,\sigma)$，利用极大似然法得到参数估计值为 $\hat{\mu}=5.262$，$\hat{\sigma}=1.112$。经计算得到免赔额为 50 元时，

每张保单的平均赔付额为 308.88 元。当免赔额提高到 100 元时，平均赔付额为 268.93 元，平均赔付额将减少 14.88%。当取消免赔额时，则每张保单的平均赔付额等于对数正态分布的期望值 356.49 元。由于医疗费用大于 50 元的概率为 0.8876，若取消免赔额，理赔次数将会增加 0.11569/0.88776 = 12.66%。当每张保单的最高赔付额不超过 2 000 元时，每张保单的平均赔付额为 287.22 元，平均赔付额将减少 7.1%。 ■

相比较而言，参数模型法对分析问题更加有用。首先，理论分布中有丰富的应用性质（如中心极限定理、独立同分布泊松随机变量的可加性），这些性质有助于对实际问题进行分析；其次，参数模型法更加简单，完全可以由少数几个参数概括，如泊松分布、指数分布只有一个参数，正态分布、对数正态分布、伽玛分布、帕累托分布和负二项式分布也仅有两个参数；最后，模型法不仅可以给出各种相关量的点估计值，还可以估计置信区间，进行误差分析。

图 1-1 是建立参数模型流程的示意图[①]，这个流程由以下六个阶段组成：

第一阶段，根据分析人员对现有数据的性质和形式的先验认知和经验，初步选择一个或多个模型。例如，在研究死亡率时，所选的模型也许会包含以下这些协变量：年龄、性别、保单生效期限、保单类型、健康方面的信息和生活方式等。在研究保险的损失量大小时，会对统计分布类型有一些自然的选择（例如：对数正态、伽玛、威布尔）。

第二阶段，基于观测数据进行模型的校准。在研究死亡率的情形，数据可能是某寿险保单群体的信息；在研究财产保险的索赔时，数据可能是某财产保险群体的实际赔付数据。

第三阶段，确定所拟合模型的有效性是否充分考虑了数据中的所有信息。这里可以采用各种诊断检验。例如一些著名的统计检验：卡方拟合优度检验、Kolmogorov - Smirnov 检验，或者按照事物的本质进行的定性检验。检验方法的选择直接依赖于建模的目的。在保险有关的研究中，经常要求最终的模型能够从整体上复制出实际经验数据所表现的损失，保险实务中也常称之为“模型的无偏性”。

第四阶段，要有适当的机会考虑其他可选模型。特别是，如果在第三阶段揭示出前面的模型不适用，这个步骤将特别有价值。在这个阶段也许会考虑不止一个的有效模型。

第五阶段，将所有第一至第四阶段考虑的有效模型按照一定的准则进

① 引自 Klugman S. A., H. H. Panjer., G. E. Willmot 著：《损失模型：从数据到决策》，吴岚译，人民邮电出版社 2009 年版，第 1~2 页。

行比较选择。这时可以利用前面获得的一些检验结果，也可以考虑其他的准则。一旦某个模型胜出，则淘汰的那些模型可用于敏感性分析。

第六阶段，最终要保证被选出的模型适用于未来的应用。也许要对参数进行适当的调整，以反映从观测数据时期到未来模型使用时期之间的通货膨胀变化。

当新的数据产生或者环境有所变化时，需要重复进行以上六个步骤以改进模型。

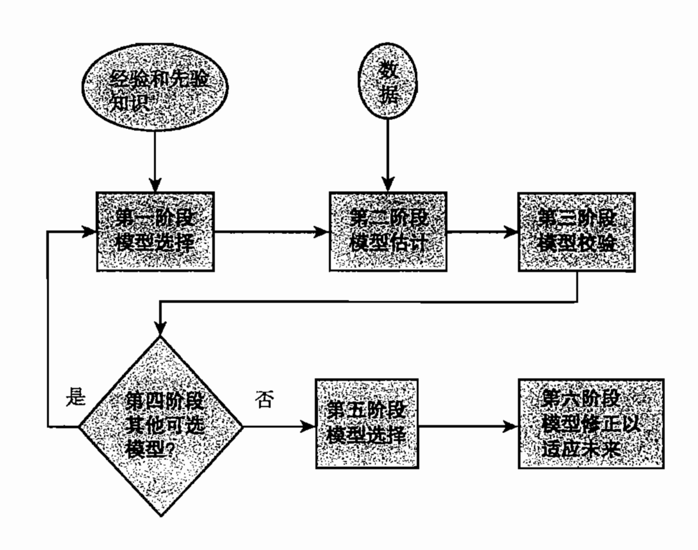


图 1－1　建模的流程

在模型选择中，要注意对模型的拟合程度和简单程度进行平衡，对模型的优劣给予客观的评价。**简洁性**是指模型中待估计的未知参数个数尽可能少，**准确性**是指模型对实际数据的拟合的误差尽可能小。另一个需要注意的问题是，对损失分布估计应考虑获得该分布的具体精算要求。在不同的精算目的下，对损失分布估计的精度要求不一样，代价也不一样。例如，如果是为了制定费率，则对损失分布的中间部分的分布情况要求较高；如果是为了考虑再保险自留额的问题，则对损失分布的尾部要作更细致的估计。

§1.2　本书的结构

本书将向读者展示整个精算建模过程。我们假定读者已经掌握了高等数学和概率统计的基础知识。我们首先要学习寿险和非寿险中的基本风险模型并介绍如何使用这些模型，然后再讨论模型参数的估计以及如何选择模型。本书结构如下：

1. 生存模型和生命表。死亡风险是寿险精算中的最主要的风险。在第

二章中将介绍寿险精算中生存模型的基本概念及形式，对生存函数进行研究，进而对死亡率和条件死亡率的概念及公式进行阐述。生命表是表达生存模型的一种最常见的形式，在第三章中将给出生命表的传统形式和标准精算符号，以之为基础，推导死亡年龄随机变量的概率密度函数，给出风险暴露数的概念和计算公式，并对非整数年龄的死亡概率进行阐述。

2. *单张保单的理赔额和理赔次数分布*。保单理赔发生的时间、发生次数和每次理赔的金额是非寿险中最主要的风险来源。在第四章中我们将介绍财产险中单张保单的理赔额和理赔次数分布。一般情况下，保险合同不可能对标的损失提供全额的赔偿。因此我们首先说明理赔额和损失额的区别与联系，其次介绍常见的保险责任：免赔额、保单限额和比例分保的基本概念，并研究带有这些保险责任后理赔额和理赔次数的分布。最后介绍几种常见的理赔次数分布的性质，并讨论免赔额对理赔次数分布的影响。

3. *短期多张保单的总理赔额的分布*。在对一些保单组合、某个业务线或者某个公司进行损失建模时，我们将关心赔付的整体情况。第五章研究个体风险模型，即保单组合中保单数已知，假设每张保单至多只发生一次理赔，且相互独立。第六章研究聚合风险模型，是既考虑赔付次数，又考虑每次赔付的金额的总理赔额模型。

4. *破产概率*。保险公司的经营过程是动态的。第七章将不仅考虑保险人长期的赔付情况，还将引入保费收入、投资收入和费用支出以及其他可能影响现金流的因素来研究保险人盈余的变化规律。我们将分别建立连续时间和离散时间盈余过程的概念和破产概率的计算公式及性质。

5. *经验模型*。有时人们需要利用数据的经验分布，这也许是因为数据的规模足够大或者是因为需要很好的表现数据本身的特点。第八章将讨论这些内容，包括对完整个体数据直接计算的简单情形、对截断或删失数据的调整、对大数据集的适当修正，特别是对生存模型的研究。

6. *参数模型的估计和选择*。构建精算模型的关键在于选择一个合适的分布模型并估计参数。第九章对不同数据类型下参数估计方法进行了详细阐述，讨论参数的区间估计和极大似然估计的方差，介绍含伴随变量的参数模型。第十章将考虑模型选择问题，包括被选模型与实际数据函数图形上的直观比较和筛选直观的选择、用统计学方法对模型分布函数进行拟合优度检验、评分法选择最优模型。

7. *估计的调整*。在此，还需要对结果作进一步的调整。本书中介绍了两种调整方法，首先是修匀调整。有时我们得到的估计值不一定满足先验的观点。例如死亡率，我们认为它应该是关于年龄的连续变化、递增的光滑曲线。而在大多数情形下，这个序列的每个元素是彼此独立得到的，它们不一定满足光滑性和递增性，所以需要采用修匀方法进行必要的调整。

第十一章讨论了表格数据修匀和参数修匀的几种方法。其次是信度调整。有时我们得到的几个估计都是基于很小的观测量得到的，数据的可信度不高，这时可以考虑添加一些相关的先验信息来提高估计的精度。第十二章的信度理论考虑了如何结合先验信息进行适当调整的机制和方法。

8. 随机模拟。在很难得到解析结果时，随机模拟（利用随机数）方法也许可以提供一些答案。第十三章讨论各种随机变量随机数的产生方法，还简单介绍了目前流行的 bootstrap 法和 MCMC 模拟法的技术。

9. 案例分析。在第十四章中，我们给出了两个案例来说明精算建模的流程。

第一篇　基本风险模型

第二章　生存分析的基本函数及生存模型

学习目标

☐ 了解对一元和多元生存模型进行分析的基本函数：生存函数、概率密度函数、危险率函数、剩余寿命均值以及剩余寿命中位数

☐ 了解五类参数生存模型：均匀分布、指数分布、Gompertz 分布、Makeham 分布以及韦伯分布

☐ 熟悉生存分析基本函数的概念及其相互关系，熟悉这些基本函数对应的精算符号

☐ 掌握五类参数生存模型的假设及结果，并能熟练运用相应的精算符号进行演算

§2.1　生存分析的基本函数

简单地说，**生存分析就是对特定事件发生的时间进行分析和推断**。由于研究领域不同，这一特定事件可以是生物体的死亡、疾病的出现、设备的失效或者债券的违约。这些事件发生的时间受随机因素的影响，是一个随机变量，通常我们称其为**生存时间随机变量**，用 T 表示，其含义是个体从初始时刻开始直至死亡、发生疾病、失效或者违约的时间，以下为了简便起见，我们称 **T 是个体从初始时刻到死亡的时间**。

T 的分布特征可以通过以下四个函数来描述：（1）反映个体存活时间超过时间 t 的概率的生存函数；（2）反映给定年龄的个体在下一瞬间死亡概率的危险率函数；（3）反映无条件瞬间死亡概率的概率密度函数；（4）反映平均死亡时间的剩余寿命均值。

这四个函数与另一个常用函数——累积危险率函数一起，刻画了随机变量 T 不同的特征。

2.1.1　生存函数

生存时间随机变量 T 表示个体从初始时刻开始的“未来寿命”，通常

我们总是从某一时刻开始记录某种生物体、某种设备的“未来寿命”，我们将这一起始时刻记为 $t=0$，在 $t=0$ 发生的事件称为**初始事件**。

描述生存时间统计特征的基本函数是生存函数，它反映被观察个体在任意时刻 $t(t\geqslant 0)$ 仍然生存的概率，我们将其定义为：

$$S(t) = P(T > t) \tag{2.1.1}$$

显然有：① $T\geqslant 0$；② $S(0) = 1$；③ $S(t)$ 是 t 的非增函数，且 $\lim_{t\to+\infty}S(t) = 0$。

生存函数有时与初始时刻个体的年龄无关，例如对已确诊患有某种重大疾病的病人而言，一旦确诊，其生存概率仅依赖于时间 t，而与病人确诊时的年龄无关。但在人寿保险与养老金计划的生存模型中，必须考虑所研究对象的确切年龄（自然年龄），这是因为不同年龄群体的生存概率 $P(T > t)$ 有很大的不同。

考虑一个 x 岁（通常 x 取整数，称其为**选择年龄**）的投保群体的生存函数，我们把保单签发的时刻作为初始时刻 $t = 0$，讨论生存函数 $S(t)$。

显然，$x = 25$ 与 $x = 45$ 所对应的函数 $S(t)$ 不一样。为了刻画 $P(T > t)$ 对 x 的依赖关系，我们通常将其记为

$$P(T > t) = S(t; x) \tag{2.1.2}$$

式（2.1.2）中的选择年龄 x 称为**伴随变量**，未来的寿命 t 称为**主要变量**。这时，生存函数为式（2.1.2）。

年龄不是唯一对生存函数 $S(t)$ 有影响的伴随变量，性别、吸烟与否等因素对未来寿命都有影响。如果考虑这些因素，那么对年龄为 x 岁的男性（m）吸烟者（s）来说，其生存函数则为 $S(t; x, m, s)$，其中 x, m, s 均为伴随变量。更一般的生存函数为 $S(t; x_1, x_2, \cdots, x_m)$。其中，$x_1, x_2, \cdots, x_m$ 是对生存有影响的 m 个伴随变量，我们称这样的生存函数为**选择生存函数**。

特殊情况下，如果在初始时刻 $x = 0$，即个体为新生婴儿，那么根据式（2.1.2），生存函数为 $S(t;0)$，也可简写为 $S(t)$。我们注意到，在时刻 t，被观察者的自然年龄也刚好是 t 岁，在不致引起混淆的情况下，可用 x 代替 t，此时新生儿的生存函数就是 $S(x)$，对应的随机变量 X 表示新生婴儿的死亡年龄（或称为**未来生命随机变量**），$S(x) = P(X > x)$。

当 T 为连续型随机变量时，生存函数与累积分布函数互补，即 $S(t) = 1 - F(t)$，这里 $F(t) = P(T\leqslant t)$，同时，生存函数也是概率密度函数 $f(t)$ 的积分，即 $S(t) = P(T > t) = \int_t^{\infty} f(y)dy$，因此，

$$f(t) = -\frac{dS(t)}{dt} \tag{2.1.3}$$

$S(t)$ 的图形叫做**生存曲线**，陡峭的生存曲线表示较低的生存概率或较短的生存时间，平缓的生存曲线表示较高的生存率或较长的生存时间。

【例 2-1】 生存时间随机变量服从指数分布，其生存函数为 $S(t) =$

$e^{-\frac{1}{\theta}t}$，$t \geq 0$，$\theta > 0$，图 2－1 为指数生存曲线。

当 T 为离散型随机变量时，假设其概率分布函数 $p(t_j) = P(T = t_j)$，$j = 1, 2, \cdots$，其中 $t_1 < t_2 < \cdots$，那么 T 的生存函数为：

$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j) \tag{2.1.4}$$

【例 2－2】　假设生存时间 T 服从离散均匀分布，概率分布函数为：

$$p(t_j) = P(T = j) = \frac{1}{3},$$

$$j = 1, 2, 3$$

其生存函数为：

$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j) = \begin{cases} 1, & 0 \leq t < 1 \\ 2/3, & 1 \leq t < 2 \\ 1/3, & 2 \leq t < 3 \\ 0, & t \geq 3 \end{cases}$$

其生存曲线如图 2－2 所示。

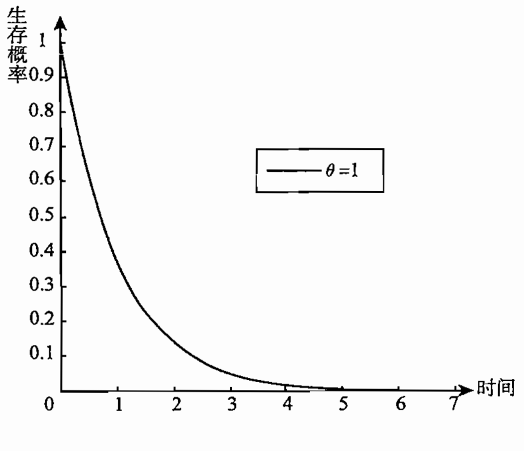


图 2－1　指数生存曲线

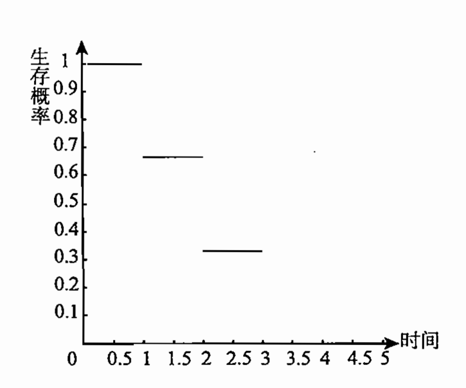


图 2－2　离散型随机寿命的生存曲线

■

2.1.2　危险率函数

危险率函数是生存分析中的另一个基本函数，它描述被观察个体在某时刻存活的条件下，在以后的单位时间内死亡的（条件）概率。危险率函数也称为**条件瞬时死亡率**、**死亡密度**。在人口学中，它也被称为**死亡力**，在可靠性研究中，也称为**条件失效率**。

危险率函数的定义为：

$$h(t) = \frac{f(t)}{S(t)} \tag{2.1.5}$$

显然，$h(t)$ 是在生存到时刻 t 的条件下的死亡密度。

我们注意到，式（2.1.3）和式（2.1.5）都是对个体在 t 时刻死亡密度的瞬时度量，但 $f(t)$ 只需个体在初始时刻生存即可，而 $h(t)$ 却需个体在时刻 t 生存，这也是称 $f(t)$ 是时刻 t 死亡的无条件密度，而 $h(t)$ 是条件密度的原因。

将式（2.1.3）代入式（2.1.5），有

$$h(t) = -\frac{dS(t)/S(t)}{dt} = -\frac{d\ln S(t)}{dt} \tag{2.1.6}$$

式（2.1.6）两边从 0 到 t 积分得：

$$\int_0^t h(y)dy = -\ln S(t) \tag{2.1.7}$$

所以，

$$S(t) = e^{-\int_0^t h(y)dy} \tag{2.1.8}$$

在实际应用中，还经常用到累积危险率函数，记为 $H(t)$，其定义为：

$$H(t) = \int_0^t h(y)dy = -\ln S(t) \tag{2.1.9}$$

则

$$S(t) = e^{-H(t)} \tag{2.1.10}$$

图 2－3 给出了常见的危险率函数曲线。

比较常见的危险率曲线是“浴盆”状危险率曲线，适合那些从出生就进入观察的人群。这种情况下，死亡率起初不断下降，主要原因在于婴儿容易患病，生命比较脆弱，之后逐渐稳定，最后随着人口的自然老化死亡率又逐渐上升。同样，对于某些设备，可能因部件损坏而在较早时候出现失效，其后一段时间危险率保持不变，到了设备的寿命后期，危险率又开始增加。如果危险率先增加，然后开始下降，则称为“驼峰”式的危险率曲线，它常用于手术成功后的生存建模。建模初期，因术后感染、出血或其他并发症等原因，危险率增加；之后随着患者的康复，危险率逐步下降。

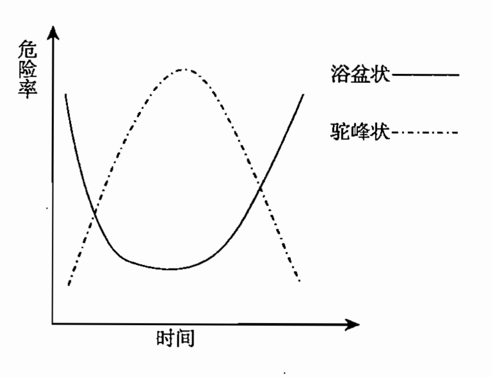


图 2－3　常见的危险率曲线

与生存函数相比，危险率函数通常可以反映死亡（失效）机制更为详细的信息。因此，在概括性地描述生存数据时，危险率函数通常占据主导地位。

到目前为止，我们已经了解了三个进行生存分析的函数：生存函数、概率密度函数以及危险率函数。

2.1.3　剩余寿命均值和剩余寿命中位数

我们要学习的第四个生存分析函数是剩余寿命均值。由于生存时间 T 通常与初始时刻个体的寿命 x 有关，因此 T 也称为**剩余寿命**。

在（0，+∞）上 $f(t)$ 可积的条件下，如果 T 连续，那么定义剩余寿命均值为 T 的期望值 $E(T)$，即

$$E(T)=\int_0^{+\infty}t\cdot f(t)dt \tag{2.1.11}$$

对（2.1.11）式进行分部积分得：

$$E(T)=\int_0^{+\infty}S(t)dt \tag{2.1.12}$$

T 的方差为：

$$Var(T)=\int_0^{+\infty}t^2\cdot f(t)dt-\left(\int_0^{+\infty}t\cdot f(t)dt\right)^2 \tag{2.1.13}$$

如果 $P(T>y)=P(T\leqslant y)=1/2$，则称 y 为随机变量 T 的中位数，也称为个体的剩余寿命中位数或剩余寿命中位数。

显然，若 y 是剩余寿命中位数，则有

$$S(y)=F(y)=\frac{1}{2} \tag{2.1.14}$$

有时，尤其是当分布的偏度较大时，中位数指标优于均值指标。

通过三个小节的内容，我们学习了反映 x 岁个体生存时间变量 T 的分布特征的四个函数。特殊地，在生命表中，新生儿的生存函数为 $S(x)$，其余三个生存分析函数及其之间的关系依然存在，只是使用的符号有所不同，如危险率函数也称为**死亡力**，用 μ_x 表示，而不是用 $h(x)$，即

$$\mu_x=-\frac{dS(x)/dx}{S(x)}=-\frac{d\ln S(x)}{dx} \tag{2.1.15}$$

习惯上用 $\mathring{e}_0$ 表示随机变量 X 的一阶矩，即

$$\mathring{e}_0=E(X)=\int_0^{+\infty}x\cdot f(x)dx \tag{2.1.16}$$

因为 $\mathring{e}_0$ 是 X 的无条件期望，因此也被称为**新生婴儿剩余寿命的完全期望**，即新生儿的平均寿命。

对于选择生存函数 $S(t;x)$，t 为随机变量的值，x 为选择年龄，那么 T 的期望值 $E(T;x)$ 给出了 x 岁人群的剩余寿命均值，用 $\mathring{e}_{[x]}$ 表示，它的危险率函数用 $\mu_{[x]+t}$ 表示，且有

$$\mu_{[x]+t}=-\frac{dS(t;x)/dt}{S(t;x)}=-\frac{d\ln S(t;x)}{dt} \tag{2.1.17}$$

§2.2　参数生存模型举例

生存模型，就是关于生存分析基本函数的数学模型。生存模型有两种

基本形式，解析形式和表格形式。前者直接给出生存分析基本函数的解析式，即参数生存模型；后者即生命表，它以表格的形式给出相关的生存分析基本函数在整数点的值。

对于被观察的个体，通常有许多原因导致其在特定的时间内死亡或失效，要想分解这些原因并从数学上准确地描述十分困难。通常，可以选择合适的理论分布来逼近生存数据。本节介绍几个可作为生存模型的非负连续型概率分布，并概括其特征，说明其应用。

2.2.1 均匀分布

均匀分布是仅有两个参数的分布，其概率密度函数为：

$$f(t) = \begin{cases} 1/(b-a), & t \in [a,b] \\ 0, & t \notin [a,b] \end{cases}$$

如果参数 $a = 0$，b 就是区间长度，同时也是 t 取到的最大值。我们常用极限年龄 ω 表示这个参数，此时其密度函数为：

$$f(t) = \begin{cases} 1/\omega, & t \in [0,\omega] \\ 0, & t \notin [0,\omega] \end{cases} \tag{2.2.1}$$

显然，式（2.2.1）表示的均匀分布有以下性质：

$$F(t) = \int_0^t f(y)dy = \frac{t}{\omega} \tag{2.2.2}$$

$$S(t) = 1 - F(t) = \int_t^\omega f(y)dy = \frac{\omega - t}{\omega} \tag{2.2.3}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{1}{\omega - t} \tag{2.2.4}$$

$$E(T) = \int_0^\omega t \cdot f(t)dt = \frac{\omega}{2} \tag{2.2.5}$$

$$Var(T) = E(T^2) - [E(T)]^2 = \frac{\omega^2}{12} \tag{2.2.6}$$

均匀分布是 Abraham de Moivre 于 1724 年提出的表示生存模型的第一个连续型概率分布。需要注意的是，在时间区间较长的情况下，将剩余寿命随机变量 T 视为均匀分布并不合适。

2.2.2 指数分布

这是一个在理论上经常使用的单参数分布，其生存函数为：

$$S(t) = e^{-\frac{1}{\theta}t}, \quad t \geq 0,\ \theta > 0 \tag{2.2.7}$$

其概率密度函数为：

$$f(t) = \frac{dS(t)}{dt} = \frac{1}{\theta}e^{-\frac{1}{\theta}t} \tag{2.2.8}$$

其危险率函数为：

$$h(t) = \frac{f(t)}{S(t)} = \frac{1}{\theta} \tag{2.2.9}$$

因为 $1/\theta$ 为常数，所以在精算教材中称其为**常值死亡力**。

指数分布的特点是历史地位重要、数学形式简单，并且具有许多重要性质。

第一，无记忆性。即

$$P(T \geqslant t + y \mid T \geqslant t) = P(T \geqslant y) \tag{2.2.10}$$

这一性质使得它在数学上易于处理，但也限制了它在可靠性研究领域的应用。

由于无记忆性，$E(T - t \mid T \geqslant t) = E(T) = \theta$，即剩余寿命均值为常数。由于未来发生某事件的时间与历史记录无关，无记忆性有时也称做“不老化”，指数分布的危险率为常数也体现了这一性质。因此，尽管历史上指数分布很受欢迎，但在健康领域和工业领域，其常数危险率的限制显得过于苛刻，不适合用做长时间段的人口生存模型，而是用于短时间区间段，如1年。

第二，指数分布的均值和标准差都是 θ，中位数是 $\theta \ln 2$。

第三，指数分布是韦伯分布和伽玛分布的特例。

由于人口生存模型很少使用均匀分布或指数分布来拟合，因此我们用 T 而不是用 X 来作为死亡时间随机变量。在下面的分布中，将用 X 表示死亡时间随机变量。

2.2.3　Gompertz 分布

Gompertz[①] 于1825年提出将该分布视为人口生存模型，其危险率定义为：

$$h(x) = Bc^x, \quad x \geqslant 0, B > 0, c > 1 \tag{2.2.11}$$

那么，生存函数为：

$$S(x) = e^{-\int_0^x h(y)dy} = \exp\left(\frac{B}{\ln c}(1 - c^x)\right) \tag{2.2.12}$$

其概率密度函数为 $h(x) \cdot S(x)$。显然其数学表达式不简洁，且分布的期望也不易求得。

2.2.4　Makeham 分布

Makeham[②] 于1860年对 Gompertz 分布进行了修正，其危险率函数为：

---

① 参见 Gompertz, B. “On the nature of the function expressive of the law of human mortality”, Phil Trans Royal Soc, London 1825, 115: 513 – 583.

② 参见 Makeham, W. M. , “On the Law of Mortality and the Construction of Annuity Tables”. J. Inst. Actuaries and Assur. Mag. 1860, 8: 301 – 310.

$$h(x) = A + Bc^x, \quad x \geqslant 0, B > 0, c > 1, A > -B \tag{2.2.13}$$

Makeham 分布假设任意年龄的危险率存在与年龄相互独立的部分，因此在 Gompertz 分布的危险率基础上加了一个常数 A。

Makeham 分布的生存函数为：

$$S(x) = \exp\left(\frac{B}{\ln c}(1 - c^x) - Ax\right) \tag{2.2.14}$$

显然，对这个分布的概率密度函数、概率分布函数、矩等数字特征进行数学处理也比较困难。

2.2.5 韦伯分布

这一分布的生存函数为：

$$S(x) = e^{-\frac{1}{\theta}x^\gamma}, \quad x \geqslant 0,\ \theta > 0,\ \gamma > 0 \tag{2.2.15}$$

这里 θ 是尺度参数或规模参数，γ 是形状参数。指数分布是韦伯分布当 $\gamma = 1$ 时的特例。图 2 - 4 给出各种韦伯生存函数的形状。

韦伯分布的危险率函数具有非常灵活的形式：

$$h(x) = \frac{\gamma}{\theta}x^{\gamma-1}, \quad x \geqslant 0,\ \theta > 0,\ \gamma > 0 \tag{2.2.16}$$

图 2 - 5 给出各种韦伯危险率曲线。

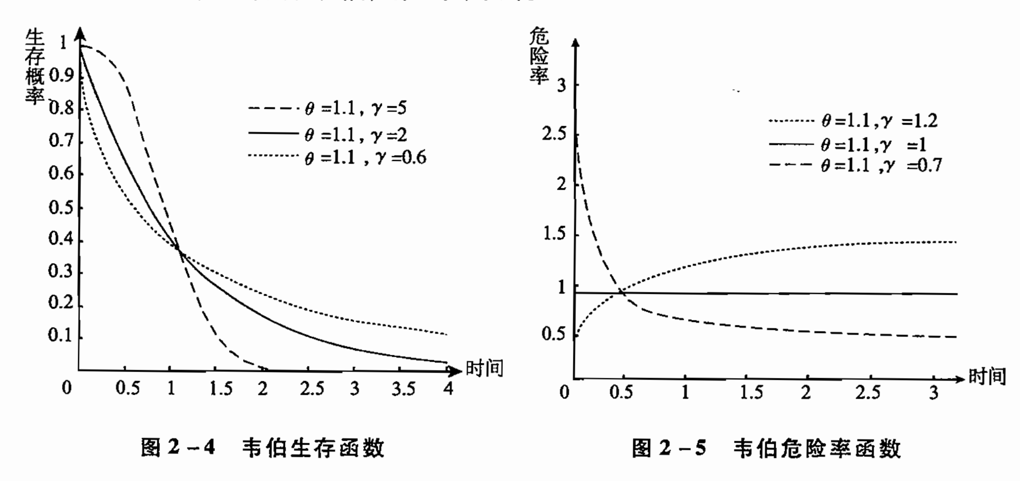


图 2 - 4　韦伯生存函数　　**图 2 - 5　韦伯危险率函数**

从图 2 - 5 可以看出，韦伯分布非常灵活，可以适用于危险率递增（取 $\gamma > 1$）、递减（取 $\gamma < 1$）和为常数（取 $\gamma = 1$）等各种情形。由于这个原因，加上韦伯分布的生存函数、危险率函数和概率密度函数形式相对简单，使得它成为应用非常广泛的参数模型。

对于韦伯分布，还有如下结果：

$$f(x) = \frac{\gamma}{\theta}x^{\gamma-1}e^{-\frac{1}{\theta}x^\gamma},\ x > 0 \tag{2.2.17}$$

$$E(X^r) = \theta^{r/\gamma}\Gamma\left(1 + \frac{r}{\gamma}\right) \tag{2.2.18}$$

$$E(X) = \theta^{1/\gamma}\Gamma\left(1 + \frac{1}{\gamma}\right) \tag{2.2.19}$$

$$Var(X) = \theta^{2/\gamma}\Gamma\left(1 + \frac{2}{\gamma}\right) - \left[\theta^{1/\gamma}\Gamma\left(1 + \frac{1}{\gamma}\right)\right]^2 \tag{2.2.20}$$

其中，$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1}e^{-t}dt$ 是著名的伽玛函数，伽玛函数的递推公式为 $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$。当 α 为整数时，$\Gamma(\alpha) = (\alpha - 1)!$；当 α 不是整数时，可见 Beyer 于 1968 年编制的伽玛函数取值表。

我们在这一节学习了五个参数生存模型。在实际运用中，没有哪个模型可以准确涵盖人的整个生命周期，我们通常可以在某个年龄段利用理论模型给出适当的拟合。

§2.3　条件随机变量的分布

在初始时刻 $t = 0$ 时，如果个体的年龄 $x = 0$，那么死亡的概率为 $F(x)$，生存的概率为 $S(x)$。这两个概率是无条件的，因为我们假定在 $x = 0$ 时所有被研究对象都活着。但构造生存模型必须考虑各个年龄的群体，因为样本可以在任何时候进入该群体。因此我们需要考虑，在被研究对象在 x 岁还活着、从年龄 x 岁开始仍然生存或死亡的概率，即条件生存函数或条件分布函数。

2.3.1　条件概率

如果某人在 x 岁生存，他在 n 年后仍生存的概率用 ${}_np_x$ 表示，那么有

$${}_np_x = \frac{S(x+n)}{S(x)} \tag{2.3.1}$$

对应的死亡概率用为 ${}_nq_x$ 表示，则

$${}_nq_x = 1 - {}_np_x = \frac{S(x) - S(x+n)}{S(x)} \tag{2.3.2}$$

这里需要注意的是条件概率 ${}_np_x$ 与非条件概率 $S(n,x)$ 的区别。二者都表达了 x 岁的人活到 $(x+n)$ 岁的概率，当已知条件是生存函数 $S(x)$ 时，它是有条件的，即 ${}_np_x = \dfrac{S(x+n)}{S(x)}$；当已知条件是选择生存函数 $S(t;x)$ 时，它是无条件的，直接由 $S(n;x)$ 求出，记为 ${}_np_{[x]}$。

同理，确定 x 岁的人在 $(x+n)$ 岁前死亡的概率时，当已知条件是生存函数 $S(x)$，它是有条件的，即 ${}_nq_x = \dfrac{S(x) - S(x+n)}{S(x)}$；当已知条件是选择生存函数 $S(t;x)$ 时，它是无条件的，直接由 $F(n;x)$ 求出，记为 ${}_nq_{[x]}$。

【例 2－3】　已知选择生存函数 $S(t;x)$，求所选取的 x 岁的人活到 $(x +$

10）岁，并在（$x+20$）岁前死亡的概率。

解：所求的概率实际上是（$x+10$）的人在 10 年内死亡的概率，即 ${}_{10}q_{[x]+10}$，那么，

$$ {}_{10}q_{[x]+10} = 1 - {}_{10}p_{[x]+10} = 1 - \frac{S(20;x)}{S(10;x)} $$ ■

2.3.2 X 的下截尾分布

对于初始时刻年龄既定的个体，设其年龄为 y，在对其进行生存分析时，需要考虑大于 y 的随机变量 X 的分布，称之为**在 y 处下截尾的 X 的分布**。

当 $x>y$ 时，这个条件变量的生存函数为：

$$ S(x|X>y) = P(X>x|X>y) = \frac{P(X>x)}{P(X>y)} = \frac{S(x)}{S(y)} \quad (2.3.3) $$

分布函数为：

$$ F(x|X>y) = P(X<x|X>y) = \frac{P(y<X<x)}{P(X>y)} = \frac{F(x)-F(y)}{1-F(y)} \quad (2.3.4) $$

特殊地，有

$$ S(y+n|X>y) = P(X>y+n|X>y) = {}_np_y \quad (2.3.5) $$

$$ F(x+n|X>x) = P(x<X\leqslant x+n|X>x) = {}_nq_x \quad (2.3.6) $$

式（2.3.3）和（2.3.4）给出了 X 在 y 处下截尾的条件变量的生存函数和累积分布函数，下面考虑条件概率密度函数。我们用 $f(x|X>y)$ 表示 y 岁的人在 x 岁死亡的条件密度，那么，

$$ f(x|X>y) = \frac{d}{dx}F(x|X>y) = \frac{d}{dx}\left[\frac{F(x)-F(y)}{1-F(y)}\right] = \frac{f(x)}{1-F(y)} $$

即
$$ f(x|X>y) = \frac{f(x)}{S(y)} \quad (2.3.7) $$

最后，对于 y 岁时生存的个体，他在 x 岁时的危险率函数 $h(x|X>y)$，有

$$ h(x|X>y) = \frac{f(x|X>y)}{S(x|X>y)} = \frac{f(x)/S(y)}{S(x)/S(y)} = \frac{f(x)}{S(x)} = h(x) \quad (2.3.8) $$

总之，$S(x|X>y)$、$F(x|X>y)$ 和 $f(x|X>y)$ 都是描述 X 在 y 处下截尾随机变量的基本函数，至于 $h(x|X>y)=h(x)$，则完全是由于危险率函数的定义所致。

2.3.3 X 的双截尾分布

我们再来考察一种更为普遍的情况，即初始时刻 y 岁的个体，在其死亡年龄不超过 z 的条件下，对其进行生存分析。也就是说，需要考虑在 X

落在 y 与 z 之间的条件下，X 的分布，称之为**在 y 和 z 处双结尾的 X 的分布**。这时，双截尾随机变量的生存函数 $S(x \mid y < X \leqslant z)$ 表示当死亡年龄 X 落在 y 与 z 之间，死亡在 x 岁以后发生的概率，有

$$\begin{aligned} S(x \mid y < X \leqslant z) &= P(X > x \mid y < X \leqslant z) \\ &= P(x < X \leqslant z \mid y < X \leqslant z) \\ &= \frac{S(x) - S(z)}{S(y) - S(z)} \end{aligned} \tag{2.3.9}$$

相应的分布函数为：

$$\begin{aligned} F(x \mid y < X \leqslant z) &= P(y < X \leqslant x \mid y < X \leqslant z) \\ &= \frac{F(x) - F(y)}{F(z) - F(y)} \end{aligned} \tag{2.3.10}$$

概率密度函数为：

$$\begin{aligned} f(x \mid y < X \leqslant z) &= -\frac{d}{dx} S(x \mid y < X \leqslant z) \\ &= -\frac{d}{dx}\left[\frac{S(x) - S(z)}{S(y) - S(z)}\right] \\ &= \frac{f(x)}{S(y) - S(z)} \end{aligned} \tag{2.3.11}$$

危险率函数为：

$$\begin{aligned} h(x \mid y < X \leqslant z) &= \frac{f(x \mid y < X \leqslant z)}{S(x \mid y < X \leqslant z)} \\ &= \frac{f(x)/[S(y) - S(z)]}{[S(x) - S(z)]/[S(y) - S(z)]} \\ &= \frac{f(x)}{S(x) - S(z)} \end{aligned} \tag{2.3.12}$$

再把 $f(x) = h(x) \cdot S(x)$ 代入式（2.3.12），有

$$h(x \mid y < X \leqslant z) = \frac{h(x)S(x)}{S(x) - S(z)} \tag{2.3.13}$$

比较式（2.3.8）与式（2.3.13），可以看出，截下尾不影响危险率函数，而截上尾却对其有影响。直观来看，截上尾的影响在于个体未来生存的时间缩短了。由式（2.3.13）可知，当 $z \to x$ 时，$h(x \mid y < X \leqslant z) \to +\infty$。

2.3.4　截尾分布的矩

X 的双截尾分布的一阶矩为：

$$E(X \mid y < X \leqslant z) = \int_y^z x \cdot f(x \mid y < X \leqslant z)\,dx \tag{2.3.14}$$

特别地，

$$E(X \mid X > y) = \int_y^{+\infty} x \cdot f(x \mid X > y)\,dx \tag{2.3.15}$$

只要上面的积分存在。

由于式（2.3.15）表示 y 岁的人死亡时的平均年龄，因此 $E(X|X>y)-y$ 就表示这个人的剩余寿命均值，用 $\mathring{e}_y$ 表示，其含义与2.1.3中的 $\mathring{e}_{[x]}$ 相似，只不过 $\mathring{e}_{[x]}$ 强调选择年龄为 x。

这里有

$$\mathring{e}_y = E(X|X>y) - y \tag{2.3.16}$$

又因为 $\int_y^{+\infty} f(x|X>y)dx = 1$，则

$$\begin{aligned}\mathring{e}_y &= \int_y^{+\infty} (x-y)f(x|X>y)dx \\ &= \int_0^{+\infty} t\cdot f(t+y|X>y)dt \end{aligned}\tag{2.3.17}$$

进一步，如果二阶矩存在，则

$$E(X^2|X>y) = \int_y^{+\infty} x^2\cdot f(x|X>y)dx \tag{2.3.18}$$

从而 y 岁人未来寿命的方差为：

$$Var(X-y|X>y) = Var(X|X>Y) = E(X^2|X>y) - [E(X|X>y)]^2 \tag{2.3.19}$$

2.3.5 中心死亡率

另一种在年龄区间 $(x,x+1]$ 上对死亡率的条件度量称为中心死亡率，用 m_x 表示。它定义为单位区间上危险率函数的加权平均值，权重是生存函数，有

$$m_x = \frac{\int_x^{x+1} S(y)h(y)dy}{\int_x^{x+1} S(y)dy} \tag{2.3.20}$$

更一般地，我们用 ${}_nm_x$ 表示区间 $(x,x+n]$ 上的中心死亡率：

$${}_nm_x = \frac{\int_x^{x+n} S(y)h(y)dy}{\int_x^{x+n} S(y)dy} \tag{2.3.21}$$

我们将在下一章对 ${}_nm_x$ 进行详细讨论，并利用它来估计表格生存函数。

【例2-4】 证明：若 X 服从指数分布，则 $m_x = -\ln p_x$。

证明： 因为 X 服从指数分布，则危险率函数 $h(y) = 1/\theta$。

由式（2.3.20），有

$$m_x = \frac{\frac{1}{\theta}\cdot\int_x^{x+1} S(y)dy}{\int_x^{x+1} S(y)dy} = \frac{1}{\theta}$$

再由 $p_x = \frac{S(x+1)}{S(x)} = \frac{e^{-\frac{1}{\theta}(x+1)}}{e^{-\frac{1}{\theta}x}} = e^{-\frac{1}{\theta}}$，可得 $\frac{1}{\theta} = -\ln p_x$，所以 $m_x = -\ln p_x$。

■

§2.4 多元生存模型

前面三节我们在对生存时间随机变量 T 分析时，仅考虑了死亡这一决定未来生存时间的因素，在实际中，往往需要考虑多种影响个体未来生存或存续时间的因素，比如解约、退休、残疾及期满等。我们把只考虑死亡这一影响因素的生存模型称为**一元生存模型**、**一元衰减模型**或者**单减因生存模型**，把考虑多种影响个体未来存续时间因素的模型称为**多元生存模型**、**多元衰减模型**或者**多减因模型**。在多元生存模型中，把个体未来生存时间称为**存续时间**，把个体由于死亡、解约、退休、残疾及期满等因素造成退出称为**个体终止**或**个体衰减**，同时，仍设个体的初始年龄为 x。

2.4.1 存续时间与减因的联合分布与边缘分布

我们仍将个体的未来存续时间记为 T，表示个体从初始时刻 $t=0$ 开始直至终止的时间；导致个体终止的衰减因素或减因记为 J，J 是离散随机变量。

假设随机变量 T 与 J 的联合概率密度函数为 $f(t,j)$，$t\geq 0$，$j=1,2,\cdots,m$，个体由于减因 j 在时刻 t 前终止的概率为 $P(T\leq t, J=j)$，记做 ${}_tq_x^{(j)}$，有

$$ {}_tq_x^{(j)} = P(T\leq t, J=j) = \int_0^t f(s,j)\,ds,\ t\geq 0,\qquad j=1,2,\cdots,m \tag{2.4.1}$$

同理，把个体由于减因 j 终止的概率记为 ${}_\infty q_x^{(j)}$，则有

$$ {}_\infty q_x^{(j)} = P(T\leq \infty, J=j) = \int_0^\infty f(s,j)\,ds,\qquad j=1,2,\cdots,m \tag{2.4.2}$$

式（2.4.2）就是随机变量 J 的边缘分布函数，记为 $r(j)$，即

$$ r(j) = {}_\infty q_x^{(j)} = \int_0^\infty f(s,j)\,ds,\qquad j=1,2,\cdots,m \tag{2.4.3}$$

个体在时刻 t 前终止的概率，即不论何种因素终止的概率记为 ${}_tq_x^{(\tau)}$，有

$$ {}_tq_x^{(\tau)} = P(T\leq t) = \sum_{j=1}^m \int_0^t f(s,j)\,ds \tag{2.4.4}$$

式（2.4.4）就是随机变量 T 的边缘分布函数，进一步，把个体在时刻 t 仍存续的概率记为 ${}_tp_x^{(\tau)}$，称之为**存续函数**，有

$$ {}_tp_x^{(\tau)} = 1 - {}_tq_x^{(\tau)} \tag{2.4.5}$$

事实上，若把 T 的边缘密度函数记为 $g(t)$ ，则 $g(t) = \sum_{j=1}^{m} f(t,j)$ ，把 T 的边缘分布函数记为 $G(t)$ ，那么，

$$G(t) = \int_0^t g(s)ds = \int_0^t \sum_{j=1}^{m} f(s,j)ds = \sum_{j=1}^{m} \int_0^t f(s,j)ds = {}_tq_x^{(\tau)} \tag{2.4.6}$$

即个体在时刻 t 前终止的概率，得到了与式（2.4.4）相同的结果。

2.4.2 危险率函数

类似于单减因生存模型，在多减因生存模型中，我们把个体在 $x+t$ 岁的危险率函数定义为：

$$h_{x+t}^{(\tau)} = \frac{G'(t)}{1-G(t)} = \frac{g(t)}{1-G(t)} \tag{2.4.7}$$

$h_{x+t}^{(\tau)}$ 表示在 x 岁的个体在存续了时间 t 后的终止密度。

由式（2.4.5）和式（2.4.4），有

$$h_{x+t}^{(\tau)} = \frac{1}{{}_tp_x^{(\tau)}} \cdot \frac{d}{dt}\,{}_tq_x^{(\tau)} = -\frac{1}{{}_tp_x^{(\tau)}} \cdot \frac{d}{dt}\,{}_tp_x^{(\tau)} = -\frac{d}{dt}\ln {}_tp_x^{(\tau)} \tag{2.4.8}$$

由（2.4.8）式可解得：

$${}_tp_x^{(\tau)} = e^{-\int_0^t h_{x+s}^{(\tau)}ds} \tag{2.4.9}$$

我们再把由衰减因素 j 导致个体在 $x+t$ 岁终止的危险率函数定义为：

$$h_{x+t}^{(j)} = \frac{f(t,j)}{1-G(t)} \tag{2.4.10}$$

自然地，我们有

$$h_{x+t}^{(j)} = \frac{f(t,j)}{{}_tp_x^{(\tau)}} = \frac{d({}_tq_x^{(j)})/dt}{{}_tp_x^{(\tau)}} \tag{2.4.11}$$

又由定义，有

$${}_tq_x^{(\tau)} = \sum_{j=1}^{m} {}_tq_x^{(j)} \tag{2.4.12}$$

把式（2.4.12）两边对 t 求导并除以 ${}_tp_x^{(\tau)}$ ，得到

$$h_{x+t}^{(\tau)} = \sum_{j=1}^{m} h_{x+t}^{(j)} \tag{2.4.13}$$

式（2.4.12）和（2.4.13）表明，由各种减因导致个体终止的概率和危险率函数都具有可加性。

根据由减因 j 导致个体在 $x+t$ 岁终止的危险率函数的定义，有

$$f(t,j) = {}_tp_x^{(\tau)} \cdot h_{x+t}^{(j)} \tag{2.4.14}$$

再由 T 的边缘密度函数 $g(t)$ 的定义可得：

$$g(t) = \sum_{j=1}^{m} f(t,j) = \sum_{j=1}^{m} {}_tp_x^{(\tau)} \cdot h_{x+t}^{(j)} = {}_tp_x^{(\tau)} \sum_{j=1}^{m} h_{x+t}^{(j)} = {}_tp_x^{(\tau)} h_{x+t}^{(\tau)} \tag{2.4.15}$$

个体在 t 时刻终止的条件下，J 的条件密度函数为：

$$h(j \mid T = t) = \frac{f(t, j)}{g(t)} = \frac{{}_tp_x^{(\tau)} h_{x+t}^{(j)}}{{}_tp_x^{(\tau)} h_{x+t}^{(\tau)}} = \frac{h_{x+t}^{(j)}}{h_{x+t}^{(\tau)}}, \qquad j = 1, 2, \cdots, m \tag{2.4.16}$$

至此，式（2.4.14）、（2.4.15）、（2.4.12）和（2.4.16）把 T 和 J 的联合概率密度函数、边缘密度函数、边缘分布函数和条件分布函数都用精算符号表示出来。

2.4.3　联合单减因生存模型

在单减因生存模型中，我们仅考虑死亡风险导致个体在一定时间内终止的概率；而在多减因生存模型中，由于各种衰减因素同时发生作用，因此某种衰减因素导致状态终止的可能性会因其他减因的存在而改变，我们称这些影响状态终止且相互作用的衰减因素为**竞争性减因**。在竞争性减因的条件下，我们只能看到这些减因对个体终止产生的总作用，而很难看到某个减因的单独作用。为了考虑各种减因的单独作用对存续函数${}_tp_x^{(\tau)}$ 和终止概率${}_tq_x^{(\tau)}$ 所造成的影响，我们可以就特定的减因定义单减因生存模型，该模型只依赖这个特定的减因，多个单减因的总和称为**联合单减因模型**。

在考虑第 j 个减因的单独作用时，相应的危险率函数 $h_{x+t}^{(j)}$ 是最基本的因素。在危险率函数 $h_{x+t}^{(j)}$ 的基础上，我们定义其他函数。

首先定义联合单减因模型的存续函数为：

$${}_tp_x^{'(j)} = e^{-\int_0^t h_{x+s}^{(j)} ds} \tag{2.4.17}$$

它表示个体在 t 时刻前仅受第 j 个减因影响所致的存续概率。再定义

$${}_tq_x^{'(j)} = 1 - {}_tp_x^{'(j)} \tag{2.4.18}$$

称为**独立终止率**。自然地，它表示在没有其他减因影响时，由第 j 个减因导致个体在 t 时刻终止的概率。“独立”一词的使用，旨在强调在确定${}_tq_x^{'(j)}$ 时，只考虑减因 j 的作用而不考虑其他减因的影响。为了区别在多减因生存模型中由减因 j 导致个体在 t 时刻前终止的概率${}_tq_x^{(j)}$，我们对${}_tq_x^{'(j)}$ 使用终止率而不用终止概率。

关于 $q_x^{(j)}$、$q_x^{'(j)}$ 和 $q_x^{(\tau)}$ 的大小，有如下关系：

$$q_x^{(j)} \leqslant q_x^{'(j)} \leqslant q_x^{(\tau)} \tag{2.4.19}$$

我们先来考虑第一个不等式。直观上来看，单纯由第 j 个减因导致个体在一年内终止的独立终止率 $q_x^{'(j)}$ 会因其他减因的作用（先于减因 j 发生）而减小。而事实上，由式（2.4.9）和（2.4.17）可知：

$${}_tp_x^{(\tau)} = e^{-\int_0^t h_{x+s}^{(\tau)} ds} = e^{-\int_0^t [h_{x+s}^{(1)} + h_{x+s}^{(2)} + \cdots + h_{x+s}^{(m)}] ds} = \prod_{j=1}^{m} {}_tp_x^{'(j)} \tag{2.4.20}$$

因此，${}_tp_x^{'(j)} \geqslant {}_tp_x^{(\tau)}$，加之 $h_{x+t}^{(j)} \geqslant 0$，所以，

$$ {}_tp_x^{'(j)} h_{x+t}^{(j)} \geqslant {}_tp_x^{(\tau)} h_{x+t}^{(j)} \tag{2.4.21} $$

式（2.4.21）两端从 0 到 1 积分便得：

$$ q_x^{'(j)} = \int_0^1 {}_tp_x^{'(j)} h_{x+t}^{(j)} dt \geqslant \int_0^1 {}_tp_x^{(\tau)} h_{x+t}^{(j)} dt = q_x^{(j)} $$

另一方面，我们又有

$$ q_x^{'(j)} = 1 - p_x^{'(j)} \leqslant 1 - p_x^{(\tau)} = q_x^{(\tau)} $$

这说明个体的独立终止率会因为减因 j 之外还有其他减因的作用而增大至 $q_x^{(\tau)}$。

对应于一元生存模型中的中心死亡率，在多元衰减模型中，我们定义全中心终止率为：

$$ m_x^{(\tau)} = \frac{\int_0^1 {}_tp_x^{(\tau)} h_{x+t}^{(\tau)} dt}{\int_0^1 {}_tp_x^{(\tau)} dt} \tag{2.4.22} $$

它是危险率函数 $h_{x+t}^{(\tau)}$ 在 x 到 $x+1$ 上的加权平均值；对应于减因 j 的中心终止率定义为：

$$ m_x^{(j)} = \frac{\int_0^1 {}_tp_x^{(\tau)} h_{x+t}^{(j)} dt}{\int_0^1 {}_tp_x^{(\tau)} dt} \tag{2.4.23} $$

它是危险率函数 $h_{x+t}^{(j)}$ 在 x 到 $x+1$ 上的加权平均值。显然，

$$ m_x^{(\tau)} = \sum_{j=1}^{m} m_x^{(j)} \tag{2.4.24} $$

在联合单减因模型中，定义独立中心终止率为：

$$ m_x^{'(j)} = \frac{\int_0^1 {}_tp_x^{'(j)} h_{x+t}^{(j)} dt}{\int_0^1 {}_tp_x^{'(j)} dt} \tag{2.4.25} $$

它是危险率函数 $h_{x+t}^{(j)}$ 在 x 到 $x+1$ 上的加权平均值，但权数是 ${}_tp_x^{'(j)}$ 而不是 ${}_tp_x^{(\tau)}$。

【例 2－5】 在一个二元衰减模型中，已知 22 岁的个体 2 年后的存续概率为 0.39，24 岁的个体 1 年内由减因 1 所致终止的概率为 0.45，3 年内由减因 2 所致终止的概率为 0.52，计算：

（1）22 岁的个体 24 岁时在 1 年内由减因 1 所致终止的概率；

（2）他 3 年内由减因 2 所致终止的概率。

解：（1）$P = {}_2p_{22}^{(\tau)} q_{24}^{(1)} = 0.39 \times 0.45 = 0.1755$

（2）$P = {}_2p_{22}^{(\tau)} {}_3q_{24}^{(2)} = 0.39 \times 0.52 = 0.2028$ ■

2.4.4 特殊假设下多减因生存模型函数与联合单减因生存模型函数的转换

我们首先考虑由多减因生存模型的生存分析函数转换成联合单减因模型相应的函数。

由于这种转换涉及个体在相邻整数年间终止分布的假设，因此我们把这里用到的假设以及相应的结论列在这里，其具体的证明参见 §3.2。

在危险率函数恒定的假设下，有 $h_x^{(\tau)} = -\ln p_x^{'(\tau)}$ ，$h_x^{(j)} = -\ln p_x^{'(j)}$ ，$j = 1, 2, \cdots, m$;

在终止服从均匀分布的假设下，有 ${}_tq_x^{(\tau)} = tq_x^{(\tau)}$, ${}_tq_x^{(j)} = tq_x^{(j)}$, $0 \leqslant t < 1$ ，$j = 1, 2, \cdots, m$ ，$f(t, j) = {}_tp_x^{(\tau)} \cdot h_{x+t}^{(j)} = q_x^{(j)}$, $j = 1, 2, \cdots, m$ 。

在多减因模型中，先假设对应于各减因的危险率函数在各年龄区间内均为常数，即

$$h_{x+t}^{(j)} = h_x^{(j)}, \qquad 0 \leqslant t < 1 ; j = 1, 2, \cdots, m \tag{2.4.26}$$

这时也有

$$h_{x+t}^{(\tau)} = h_x^{(\tau)}, \qquad 0 \leqslant t < 1 \tag{2.4.27}$$

于是可得：

$$\begin{aligned} q_x^{(j)} &= \int_0^1 {}_tp_x^{(\tau)} h_x^{(j)} dt = \frac{h_x^{(j)}}{h_x^{(\tau)}} \int_0^1 {}_tp_x^{(\tau)} h_x^{(\tau)} dt \\ &= \frac{\ln p_x^{'(j)}}{\ln p_x^{(\tau)}} q_x^{(\tau)} \end{aligned} \tag{2.4.28}$$

由式（2.4.28）可解得：

$$q_x^{'(j)} = 1 - (1 - q_x^{(\tau)})^{\frac{q_x^{(j)}}{q_x^{(\tau)}}} \tag{2.4.29}$$

再假设各种终止事件的发生在各年龄区间服从均匀分布，即

$${}_tq_x^{(j)} = tq_x^{(j)}, \qquad 0 \leqslant t < 1 ; j = 1, 2, \cdots, m \tag{2.4.30}$$

因此，

$${}_tq_x^{(\tau)} = tq_x^{(\tau)}, \qquad 0 \leqslant t < 1 \tag{2.4.31}$$

由式（2.4.12），有

$$\begin{aligned} h_{x+t}^{(j)} &= \frac{1}{{}_tp_x^{(\tau)}} \frac{d}{dt} {}_tq_x^{(j)} = \frac{1}{1 - tq_x^{(\tau)}} \frac{d}{dt}(tq_x^{(j)}) \\ &= \frac{q_x^{(j)}}{1 - tq_x^{(\tau)}} \end{aligned} \tag{2.4.32}$$

所以，

$$q_x^{'(j)} = 1 - e^{-\int_0^1 h_{x+t}^{(j)} dt} = 1 - e^{-\int_0^1 \frac{q_x^{(j)}}{1 - tq_x^{(\tau)}} dt} = 1 - e^{\frac{q_x^{(j)}}{q_x^{(\tau)}} \ln(1 - q_x^{(\tau)})} = 1 - (1 - q_x^{(\tau)})^{\frac{q_x^{(j)}}{q_x^{(\tau)}}}$$

即

$$q_x^{'(j)} = 1 - (1 - q_x^{(\tau)})^{\frac{q_x^{(j)}}{q_x^{(\tau)}}} \tag{2.4.33}$$

比较式（2.4.29）和（2.4.33），可见在这两种假设下，根据终止概率算

得的独立终止率是相同的。

我们接着考虑由联合单减因模型的生存分析函数转换成多元衰减模型的函数。在多元衰减模型中，无论危险率函数在各年龄区间内均为常数，或各减因在各年龄区间服从均匀分布，都有式（2.4.28），把式（2.4.20）代入式（2.4.28），有

$$q_x^{(j)} = \frac{\ln p_x^{'(j)}}{\ln \prod_{j=1}^{m} p_x^{'(j)}}\left(1 - \prod_{j=1}^{m} p_x^{'(j)}\right)$$

$$= \frac{\ln\left(1 - q_x^{'(j)}\right)}{\sum_{j=1}^{m} \ln\left(1 - q_x^{'(j)}\right)}\left[1 - \prod_{j=1}^{m}\left(1 - q_x^{'(j)}\right)\right] \tag{2.4.34}$$

由式（2.4.34）可以根据独立终止率计算终止概率。

但是，式（2.4.34）要求 $p_x^{'(j)} \neq 0 (j = 1, 2, \cdots, m)$。当 $p_x^{'(j)} = 0$ 或 $p_x^{(\tau)} = 0$ 时，就需要考虑其他方法。

一种适合处理这种不确定性的方法是改变假设，比如假设联合单减因模型中终止事件的发生在各年龄区间内服从均匀分布，而不是如上所假设的多元风险模型中各终止事件的发生在各年龄内服从均匀分布，这样，有

$${}_tq_x^{'(j)} = tq_x^{'(j)}, \qquad 0 \leqslant t < 1; j = 1, 2, \cdots, m$$

从而

$${}_tp_x^{'(j)} h_{x+t}^{(j)} = \frac{d\left({}_tq_x^{'(j)}\right)}{dt} = \frac{d\left(tq_x^{'(j)}\right)}{dt} = q_x^{'(j)} \tag{2.4.35}$$

于是有

$$q_x^{(j)} = \int_0^1 {}_tp_x^{(\tau)} h_{x+t}^{(j)} dt = \int_0^1 {}_tp_x^{'(j)} h_{x+t}^{(j)} \prod_{i \neq j} {}_tp_x^{'(i)} dt$$

$$= q_x^{'(j)} \int_0^1 \prod_{i \neq j}\left(1 - tq_x^{'(i)}\right) dt \tag{2.4.36}$$

式（2.4.36）中被积函数是一个关于 t 的多项式，可以直接进行积分运算，但在 m 较大时计算比较繁琐。

比较式（2.4.29）、（2.4.34）和（2.4.36），无论是计算 $q_x^{'(j)}$，还是计算 $q_x^{(j)}$，运算都比较复杂，因此有必要寻求近似的方法来简化计算。

第一，在多减因模型中，当对应于各减因的危险率函数在各年龄区间内均为常数时，有

$$m_x^{(j)} = \frac{\int_0^1 {}_tp_x^{(\tau)} h_{x+t}^{(j)} dt}{\int_0^1 {}_tp_x^{(\tau)} dt} = \frac{\int_0^1 {}_tp_x^{(\tau)} h_x^{(j)} dt}{\int_0^1 {}_tp_x^{(\tau)} dt} = h_x^{(j)} \tag{2.4.37}$$

同理也有

$$m_x^{'(j)} = \frac{\int_0^1 {}_tp_x^{'(j)} h_{x+t}^{(j)} dt}{\int_0^1 {}_tp_x^{'(j)} dt} = h_x^{(j)} \tag{2.4.38}$$

因此有

$$m_x^{(j)} = m_x^{'(j)} \tag{2.4.39}$$

一般情况下，有

$$m_x^{(j)} \approx m_x^{'(j)} \tag{2.4.40}$$

第二，若各终止事件的发生在各年龄区间内服从均匀分布，则

$$m_x^{(j)} = \frac{\int_0^1 {}_tp_x^{(\tau)} h_{x+t}^{(j)} dt}{\int_0^1 {}_tp_x^{(\tau)} dt} = \frac{q_x^{(j)}}{\int_0^1 (1 - tq_x^{(\tau)}) dt} = \frac{q_x^{(j)}}{1 - \frac{1}{2} q_x^{(\tau)}} \tag{2.4.41}$$

第三，若各联合单减因模型中终止事件的发生在各年龄区间内服从均匀分布，则

$$m_x^{'(j)} = \frac{\int_0^1 {}_tp_x^{'(j)} h_{x+t}^{(j)} dt}{\int_0^1 {}_tp_x^{'(j)} dt} = \frac{q_x^{'(j)}}{\int_0^1 (1 - tq_x^{'(j)}) dt} = \frac{q_x^{'(j)}}{1 - \frac{1}{2} q_x^{'(j)}} \tag{2.4.42}$$

事实上，第二个假设与第三个假设是有差异的，因为第二个假设意味着 ${}_tp_x^{(\tau)} = 1 - tq_x^{(\tau)}$，而第三个假设意味着 ${}_tp_x^{(\tau)} = {}_tp_x^{'(1)}\ {}_tp_x^{'(2)}\ {}_tp_x^{'(3)} = (1 - tq_x^{'(1)})(1 - tq_x^{'(2)})(1 - tq_x^{'(3)})$，这说明这两个假设并不等价，但是出于近似的考虑，我们仍可令

$$\frac{q_x^{(j)}}{1 - \frac{1}{2} q_x^{(\tau)}} \approx \frac{q_x^{'(j)}}{1 - \frac{1}{2} q_x^{'(j)}} \tag{2.4.43}$$

分别就 $q_x^{'(j)}$ 和 $q_x^{(j)}$ 解得：

$$q_x^{'(j)} \approx \frac{q_x^{(j)}}{1 - \frac{1}{2}(q_x^{(\tau)} - q_x^{(j)})} \tag{2.4.44}$$

$$q_x^{(j)} \approx \frac{q_x^{'(j)}(1 - \frac{1}{2} q_x^{(\tau)})}{1 - \frac{1}{2} q_x^{'(j)}} \tag{2.4.45}$$

式（2.4.44）和（2.4.45）只涉及简单的四则运算。

【例 2-6】 在有两个减因的模型中，已知 $m_{40}^{(\tau)} = 0.2, q_{40}^{'(1)} = 0.1$，在下列假设下计算 $q_{40}^{'(2)}$：

（1）多减因生存模型中，各个终止事件的发生在各年龄区间内服从均匀分布；

（2）联合单减因生存模型中，各个终止事件的发生在各年龄区间内服从均匀分布。

解：

（1）由假设可知 $m_{40}^{(\tau)} = \dfrac{q_{40}^{(\tau)}}{1 - \frac{1}{2} q_{40}^{(\tau)}} = 0.2$，得 $q_{40}^{(\tau)} = \dfrac{2}{11}$

由 $q_{40}^{'(1)}=1-(1-q_{40}^{(\tau)})^{\frac{q_{40}^{(1)}}{q_{40}^{(\tau)}}}$ 得 $q_{40}^{(1)}=0.09546$

再由 $q_{40}^{(\tau)}=q_{40}^{(1)}+q_{40}^{(2)}$ 可得 $q_{40}^{(2)}=0.08636$

因此，$q_{40}^{'(2)}=1-[1-q_{40}^{(\tau)}]^{\frac{q_{40}^{(2)}}{q_{40}^{(\tau)}}}=0.090916$

（2）由已知可知 $m_{40}^{(1)}+m_{40}^{(2)}=m_{40}^{(\tau)}$，所以，

$$\frac{\int_0^1 {}_tp_{40}^{(\tau)}h_{40+t}^{(1)}dt}{\int_0^1 {}_tp_{40}^{(\tau)}dt}+\frac{\int_0^1 {}_tp_{40}^{(\tau)}h_{40+t}^{(2)}dt}{\int_0^1 {}_tp_{40}^{(\tau)}dt}=0.2$$

即

$$\frac{q_{40}^{'(1)}\int_0^1 {}_tp_{40}^{'(2)}dt}{\int_0^1 {}_tp_{40}^{'(1)}\,{}_tp_{40}^{'(2)}dt}+\frac{q_{40}^{'(2)}\int_0^1 {}_tp_{40}^{'(1)}dt}{\int_0^1 {}_tp_{40}^{'(1)}\,{}_tp_{40}^{'(2)}dt}=\frac{q_{40}^{'(1)}}{\int_0^1 {}_tp_{40}^{'(1)}dt}+\frac{q_{40}^{'(2)}}{\int_0^1 {}_tp_{40}^{'(2)}dt}$$

$$=\frac{q_{40}^{'(1)}}{1-\frac{1}{2}q_{40}^{'(1)}}+\frac{q_{40}^{'(2)}}{1-\frac{1}{2}q_{40}^{'(2)}}=0.2$$

经计算可得 $q_{40}^{'(2)}=0.09045$。 ■

习　题

1. 生存函数 $S(t)=0.1\times(100-t)^{\frac{1}{2}},0\leqslant t\leqslant 100$，求下列各值：

(1) $f(36)$；(2) $h(50)$；(3) $H(75)$；(4) $E(T)$；(5) $Var(T)$。

2. 假设生存函数为 $S(x)=ax^2+b,0\leqslant x\leqslant k$，若 X 的期望为 60，求 X 的中位数。

3. X_1 和 X_2 为相互独立的随机变量，定义随机变量 $Y=\min(X_1,X_2)$ 与 $Z=\max(X_1,X_2)$。

(1) 证明 Y 的生存函数为 X_1 与 X_2 的生存函数的乘积；

(2) 证明 Z 的累积分布函数为 X_1 与 X_2 的累积分布函数的乘积；

(3) 证明若 X_1、X_2 都服从指数分布，则 Y 也服从指数分布，但 Z 不服从指数分布。

4. 已知 $S(x)=\left(1-\frac{x}{100}\right)^2,0\leqslant x<100$，试求 $F(75)$，$f(75)$ 和 $h(75)$。

5. 记 ${}_{m|}q_0=S(m)-S(m+1)$ 为在初始时刻 $t=0$ 生存的个体在 $t=m$ 至 $t=m+1$ 间死亡的概率。试确定，当 T 为下列分布时 ${}_{m|}q_0$ 是 m 的增函数、减函数还是常数函数？

(1) 均匀分布；(2) 指数分布；(3) $f(t)=0.00125t,0\leqslant t\leqslant 40$

6. 令 $S(x) = ax + b, 0 \leqslant x \leqslant k$，若已知$_2 p_0 = 0.075$，求$_2 m_3$。

7. 对于某一实验设备，其生存分布为在 $t_1 = 5$ 和 $t_2 = 15$ 双截尾的指数分布，参数 $\theta = 0.02$，求该试验设备未来寿命的中位数。

8. 若 $S(x) = 1 - (0.01x)^2, 0 \leqslant x \leqslant 100$，求中位数年龄时的未来预期寿命。

9. 假设一类下截尾分布的分布函数为 $S(x) = \exp\left[-\frac{Bc^x}{\ln c}\right], -\infty < x < +\infty$

（1）证明若截去 $x = 0$ 以下的分布，该分布就是 Gompertz 分布；

（2）比较未截尾分布的危险率函数和 Gompertz 分布的危险率函数。

10. 某项调查针对 50～55 岁男子的死亡率，观察的样本数量为 n，观察时间从样本 50 岁生日开始，直到死亡或者存活到 55 岁，假定危险率函数 $h(t)$ 的形式如下：

$h(t) = \alpha + \beta t$（$\alpha > 0, \beta > 0$ 是要估计的参数，t 是 50 岁以后的生存整年）

（1）推导 50～55 岁的生存函数；

（2）画出该生存函数图，并评价危险率函数在这个年龄范围上的合理性。

11. 已知某两减因生存模型满足以下条件：（1）$q_x^{'(2)} = 2q_x^{'(1)}$；（2）$q_x^{'(1)} + q_x^{'(2)} = q_x^{(\tau)} + 0.18$。试计算 $q_x^{'(2)}$。

12. 假设 $q_x^{'(1)} = 0.15$，$q_x^{'(2)} = 0.03$，如果减因 1 的联合单减因表均匀分布假设，而减因 2 集中发生在每个年龄的中间，试计算：（1）$_t p_x^{(\tau)}$，$0 \leqslant t \leqslant 1$；（2）$q_x^{(1)}, q_x^{(2)}$。

13. 某公司员工李先生，年龄为 50 岁。假设这个公司员工的寿命规律服从一个双减因模型：

（1）减因 1 是退休。

（2）$h_{50+t}^{(1)} = \begin{cases} 0.00, & 0 \leqslant t \leqslant 5 \\ 0.02, & 5 \leqslant t \end{cases}$

（3）减因 2 是其他所有非退休原因，从而停止在这个公司工作。

（4）$h_{50+t}^{(2)} = \begin{cases} 0.05, & 0 \leqslant t \leqslant 5 \\ 0.03, & 5 \leqslant t \end{cases}$

（5）如果李先生停止工作并离开此公司，那么他不会重新加入该公司。计算李先生将会在 60 岁之前退休离开该公司的概率。

14. 一个双减因模型的信息如下：

（1）$\mu_x^{(1)}(t) = 0.2\mu_x^{(\tau)}(t)$，$t > 0$

（2）$\mu_x^{(\tau)}(t) = kt^2$，$t > 0$

（3）$q_x^{'(1)} = 0.04$

计算$_2 q_x^{(2)}$。

15. 某生命个体现在年龄为（X）岁：（1）K 是整数未来寿命的随机变

量；(2) $q_{x+k}=0.1(k+1)$，$k=0,1,2,\cdots,9$，计算 $Var(K\wedge 3)$。

16. 生存函数 $S(x)$ 满足以下条件：

$$S(x)=\begin{cases}1, & 0\leqslant x\leqslant 1\\ 1-\dfrac{e^x}{100}, & 1\leqslant x\leqslant 4.5\\ 0, & x\geqslant 4.5\end{cases}$$

计算 $h(4)$。

17. 一个三减因生存模型信息如下：(1) $h_x^{(1)}(t)=0.3$，$t>0$；(2) $h_x^{(2)}(t)=0.5$，$t>0$；(3) $h_x^{(3)}(t)=0.7$，$t>0$。计算 $q_x^{(2)}$。

18. 一个双减因模型中，减因 1 是死亡，减因 2 是退保，具体信息如下：(1) 联合单减因模型中死亡在每个生命年中均匀分布；(2) 退保只发生在每年年末；(3) $l_x^{(\tau)}=1\ 000$；(4) $q_x^{(2)}=0.4$；(5) $d_x^{(1)}=0.45d_x^{(2)}$。计算 $p_x'^{(2)}$。

19. 一个三减因的生存模型中，给出如下信息：(1) 在其相应的联合单减因模型中，每一个减因在每个整数年间都服从均匀分布。(2) $q_x'^{(1)}=0.200$；(3) $q_x'^{(2)}=0.080$；(4) $q_x'^{(3)}=0.125$。计算 $q_x^{(1)}$。

20. 一个 1 000 人的群体，年龄均为 60 岁，服从三个减因：死亡；残疾；退休。给出如下信息：(1) 独立终止率如表 2－1 所示；(2) 在多减因模型中各减因在整数年龄间服从均匀分布。计算在 62 岁之前退休人数的期望值。

表 2－1

x	$q_x'^{(1)}$	$q_x'^{(2)}$	$q_x'^{(3)}$
60	0.010	0.030	0.100
61	0.013	0.050	0.200

第三章 生 命 表

学习目标

- □ 了解生命表的基本内容、构造原理及种类
- □ 熟悉生命表函数与生存分析函数之间的关系，特别是不同假设下整数年龄间生命表函数的推导
- □ 掌握并运用不同的生命表解决实际问题

§3.1 生命表及其内容

3.1.1 生命表的基本内容

在§2.2 中我们指出，生命表是给出生存分析基本函数在整数点值的表，是反映在封闭人口的条件下，一批人从出生后陆续死亡的全部过程的一种统计表。所谓**封闭人口，是指所观察的一批人只有死亡变动，没有因出生而产生的新增人口和迁入或迁出人口。**

人类历史上最早出现的生命表是哈雷彗星的发现者哈雷(Halley) 在**1693 年**根据伦敦附近人口死亡的统计规律编制出来的。因为生命表在寿险精算中的重要地位，这一年也被称为精算学的诞生之年。

通常，生命表采用表 3－1 的形式编制。

表 3－1 生 命 表

x	l_x
0	1 000 000
1	996 963
2	994 813
…	…
50	
…	…
70	687 094
…	…
90	99 580
…	…
110	0

在表 3－1 中，出现了最基本的生命表函数 l_x，它表示存活到年龄 x 岁的人口数，$x=0$，1，…，$\omega-1$。在存活人数中，l_0 是 $t=0$ 时刻新生儿的人数，称为“生命表基数”，由于我们关心的是一批人在生命期的死亡规律，因此最初人口的绝对数并不重要，研究中可以任意取值，出于方便通常取 10 的整数次幂。ω 是人口生命极限年龄，是生命表的年龄上限，在表 3－1 中 $\omega=110$。

为使用方便，生命表中还有其他函数，主要有：

(1)$_nd_x$：x 岁的人在 $(x, x+n]$ 死亡的人数。即

$$_nd_x = l_x - l_{x+n} \tag{3.1.1}$$

简记$_1d_x$ 为 d_x，则

$$l_0 = \sum_{x=0}^{\omega-1} d_x \tag{3.1.2}$$

(2)$_nq_x$：x 岁的人在 $(x, x+n]$ 死亡的概率。即

$$_nq_x = \frac{_nd_x}{l_x} \tag{3.1.3}$$

简记$_1q_x$ 为 q_x。$_nq_x$ 在第二章 2.3.1 中已经出现过，其含义没有发生变化，在这一章我们把它用生命表函数表示出来。在已知 q_x 后，依生命表基数 l_0 可以计算出各年龄的存活人数和死亡人数，生命表正是以分析各整数年龄死亡概率为基础编制出来的。

与 2.3.1 中的条件概率相同，我们有$_np_x = 1 - {_nq_x}$，而且由式（3.1.1）和（3.1.3）我们还有

$$_np_x = \frac{l_{x+n}}{l_x} \tag{3.1.4}$$

简记$_1p_x$ 为 p_x。

(3)$_nL_x$：x 岁的人群在 $(x, x+n]$ 生存的人年数。即

$$_nL_x = \int_x^{x+n} l_t dt \tag{3.1.5}$$

$_nL_x$ 表示这一时间段内暴露于死亡风险下所有生存个体存活时间的总和，也称为“暴露数”，是一个表示人群存活总时间的复合单位，简记$_1L_x$ 为 L_x。

假设死亡时间在整数年龄段内均匀分布，那么 $(x, x+n]$ 上的死亡人数$_nd_x$ 平均存活 $\frac{n}{2}$ 年，活到 $(x+n)$ 岁的 l_{x+n} 个人存活 n 年，因此，

$$_nL_x \approx n \cdot l_{x+n} + \frac{n}{2} \cdot {_nd_x} = \frac{n}{2}(l_x + l_{x+n}) \tag{3.1.6}$$

特殊地，有

$$L_x \approx \frac{1}{2}(l_x + l_{x+1}) \tag{3.1.7}$$

(4) T_x：x 岁的人群在未来的累积生存人年数。即

$$T_x = \sum_{t=x}^{\omega-1} L_t = \int_x^{\omega} l_t dt \tag{3.1.8}$$

(5) $\mathring{e}_x$：x 岁人群的平均剩余寿命。这一符号在第二章的 2.1.3 和 2.3.4 都已出现过，但我们在这一章用生命表函数来表示它。即

$$\mathring{e}_x = \frac{T_x}{l_x} = \frac{\int_0^{+\infty} l_{x+t} dt}{l_x} = \int_0^{+\infty} {_tp_x} dt \tag{3.1.9}$$

我们再来考虑新生儿的平均寿命 $\mathring{e}_0$。假设死亡时间在整数年龄间均匀

分布，由式（3.1.9）和（3.1.7），有

$$\mathring{e}_0 = \frac{T_0}{l_0} = \frac{(L_0 + L_1 + L_2 + \cdots + L_{\omega-1})}{l_0}$$

$$= \frac{1}{l_0} \times \frac{1}{2}[(l_0 + l_1) + (l_1 + l_2) + \cdots + (l_{\omega-1} + l_\omega)]$$

$$= \frac{1}{l_0} \times \left[\frac{1}{2}\sum_{t=0}^{\omega-1} d_t + \sum_{t=1}^{\omega-1} d_t + \sum_{t=2}^{\omega-1} d_t + \cdots + d_{\omega-1}\right]$$

$$= \frac{1}{l_0} \times \left[\frac{1}{2}d_0 + \left(1 + \frac{1}{2}\right)d_1 + \left(2 + \frac{1}{2}\right)d_2 + \cdots + \left(\omega - 1 + \frac{1}{2}\right)d_{\omega-1}\right]$$

$$= \frac{1}{l_0}\sum_{t=0}^{\omega-1}\left[\left(t + \frac{1}{2}\right)d_t\right]$$

即

$$\mathring{e}_0 = \sum_{t=0}^{\omega-1}\left[\frac{d_t}{l_0}\left(t + \frac{1}{2}\right)\right] \tag{3.1.10}$$

下面我们来分析式（3.1.10）。我们假设死亡时间在整数年龄间均匀分布，那么 $t+0.5$ 是死亡者的平均年龄。注意到 $l_0 = \sum_{t=0}^{\omega-1} d_t$，可以看出，新生儿的平均寿命是一个以各年龄死亡人数比例为权重的平均死亡年龄。

运用生命表基本函数，我们还可以定义寿险精算中另一个常用的死亡概率——延期死亡概率。以 ${}_{n|m}q_x$ 表示 x 岁的人延期 n 年在 $(x+n, x+n+m]$ 岁之间死亡的概率，有

$${}_{n|m}q_x = \frac{{}_m d_{x+n}}{l_x} \tag{3.1.11}$$

$$= \frac{l_{x+n} - l_{x+n+m}}{l_x} = {}_n p_x - {}_{n+m} p_x$$

$$= {}_n p_x \cdot {}_m q_{x+n} \tag{3.1.12}$$

简记 ${}_{n|1}q_x$ 为 ${}_{n|}q_x$，特殊地，有 ${}_{n|0}q_x = 0$；${}_{n|\infty}q_x = {}_n p_x \cdot {}_\infty q_{x+n} = {}_n p_x$。

3.1.2　生命表函数与生存分析函数

在初始时刻由 l_0 个新生儿组成的封闭人群中，记其成员的寿命分别为 $X_1, X_2, \cdots, X_{l_0}$，$X_i > 0$，$i = 1, 2, \cdots, l_0$，假定这是一组连续且独立同分布的随机变量，与 X 同分布。

在时刻 $t = x$，记这一人群的人数为 N_x，有

$$N_x = \sum_{k=1}^{l_0} I(X_k \geqslant x) \tag{3.1.13}$$

$$l_x = E[N_x] = E\left[\sum_{k=1}^{l_0} I(X_k \geqslant x)\right] \tag{3.1.14}$$

$$= \sum_{k=1}^{l_0} E[I(X_k \geqslant x)] = \sum_{k=1}^{l_0} P(X_k \geqslant x)$$

$$= l_0 \cdot {}_x p_0 \tag{3.1.15}$$

$$ {}_{x}p_{0} = \frac{l_x}{l_0} \tag{3.1.16} $$

可见，在给定生命表基数 l_0 的前提下，生命表函数 $l_x, x \geqslant 0$ 和生存分析函数之间存在一一对应关系。所以，有了生命表，就可以计算出${}_{x}p_{0}$，$x=0$，1，2，…，相应地可以得到${}_{x}q_{0}$，$x=0$，1，2，…

我们在2.1.1中曾记新生儿的生存函数为 $S(x)$，表示新生婴儿的剩余寿命随机变量 X 大于 x 的概率，即 $S(x) = {}_{x}p_{0}$，因此通过生命表中的 l_x 就可以完全掌握 $S(x)$ 在各整数点的值。但这是不够的，我们还需要其他生存分析的函数来确定相邻整数年间的生存人数。

1. *死亡力*。在2.1.3中，我们提到把新生儿的风险率函数称为"死亡力"，用 μ_x 表示，即 $\mu_x = -\frac{dS(x)}{S(x)}$。

由于 $l_x = l_0 \cdot S(x)$，所以，

$$ \mu_x = -\frac{dl_x}{l_x} = -\frac{d\ln l_x}{dx} \tag{3.1.17} $$

类似于式（2.1.8）的推导，我们有

$$ l_x = l_0 e^{-\int_0^x \mu_y dy} \tag{3.1.18} $$

2. *X 的概率密度*。由式（2.1.5）可知 $f(x) = h(x)S(x)$，因此我们有

$$ f(x) = S(x) \cdot \mu_x = {}_{x}p_{0} \cdot \mu_x \tag{3.1.19} $$

另外，由式（2.1.15），有 $\frac{dS(x)}{dx} = -S(x) \cdot \mu_x$，即

$$ \frac{d\,{}_{x}p_{0}}{dx} = -{}_{x}p_{0} \cdot \mu_x \tag{3.1.20} $$

3. *X 的方差*。X 的二阶矩为：

$$ E(X^2) = \int_0^\infty x^2 \cdot f(x)dx = \int_0^\infty x^2 \cdot {}_{x}p_{0} \cdot \mu_x dx \tag{3.1.21} $$

对式（3.1.21）分部积分得：

$$ E(X^2) = 2\int_0^\infty x \cdot {}_{x}p_{0} dx = \frac{2\int_0^\infty x \cdot l_x dx}{l_0} \tag{3.1.22} $$

再由式（3.1.8）得：

$$ \frac{dT_x}{dx} = -l_x \tag{3.1.23} $$

利用式（3.1.23）对式（3.1.22）的分子分部积分，有 $E(X^2) = \frac{2\int_0^\infty T_x dx}{l_0}$，令

$$Y_0 = \int_0^{\infty} T_x dx \tag{3.1.24}$$

则
$$E(X^2) = \frac{2Y_0}{l_0} \tag{3.1.25}$$

因此，
$$Var(X) = \frac{2Y_0}{l_0} - \left(\frac{T_0}{l_0}\right)^2 \tag{3.1.26}$$

4. x 岁人剩余寿命的概率密度。在第二章中，我们定义了个体生存时间为 T。在生命表中，若个体在初始时刻的年龄为 x，我们记其剩余寿命为 $T(x)$，仍简记为 T，则有 $T = X - x$。由式（2.3.7）有

$$f_T(t \mid X > x) = f_X(x + t \mid X > x) = \frac{f_X(x + t)}{S(x)} \tag{3.1.27}$$

由式（3.1.19）得 $f_X(x + t) = \mu_{x+t} \frac{l_{x+t}}{l_0}$，又由 $S(x) = \frac{l_x}{l_0}$，从而，

$$f_T(t \mid X > x) = \frac{l_{x+t}}{l_x} \cdot \mu_{x+t} = {}_t p_x \cdot \mu_{x+t} \tag{3.1.28}$$

又因为 ${}_t p_x = P(T > t \mid X > x) = 1 - P(T \leqslant t \mid X > x) = 1 - F_T(t \mid X > x)$，所以，

$$\frac{\partial}{\partial t}({}_t p_x) = -\frac{\partial}{\partial t}[F_T(t \mid X > x)] = -f_T(t \mid X > x)$$

$$= -{}_t p_x \cdot \mu_{x+t} \tag{3.1.29}$$

相对随机变量 T，若把个体在初始时刻的年龄 x 当作常量，则（3.1.29）也可记为：

$$\frac{d}{dt}({}_t p_x) = -{}_t p_x \cdot \mu_{x+t} \tag{3.1.30}$$

用式（3.1.30），我们还可以得到 x 岁人平均剩余寿命 $\mathring{e}_x$ 的又一个表达式：

$$\mathring{e}_x = E(T) = \int_0^{+\infty} tf(t \mid X > x)dt = \int_0^{+\infty} t \cdot {}_t p_x \mu_{x+t} dt$$

$$= \int_0^{+\infty} {}_t p_x dt \tag{3.1.31}$$

$$= \frac{T_x}{l_x} \tag{3.1.32}$$

以及
$$E(T^2) = \frac{2Y_x}{l_x} \tag{3.1.33}$$

和
$$Var(T) = \frac{2Y_x}{l_x} - \left(\frac{T_x}{l_x}\right)^2 \tag{3.1.34}$$

其中，
$$Y_x = \int_x^{+\infty} T_y dy \tag{3.1.35}$$

5. 中心死亡率。在人口学中，还有一个更常用的死亡率概念，即在 2.3.5 中定义的中心死亡率 ${}_n m_x$。由式（2.3.21）可知：

$$
{}_n m_x = \frac{\int_0^n S(x+t)h(x+t)dt}{\int_0^n S(x+t)dt} = \frac{\int_0^n \frac{S(x+t)}{S(x)}h(x+t)dt}{\int_0^n \frac{S(x+t)}{S(x)}dt}
$$

$$
= \frac{\int_0^n {}_t p_x \cdot \mu_{x+t} dt}{\int_0^n {}_t p_x dt} = \frac{l_x \times \int_0^n {}_t p_x \cdot \mu_{x+t} dt}{l_x \times \int_0^n {}_t p_x dt}
$$

$$
= \frac{l_x \times \int_0^n f_T(t \mid X > x) dt}{l_x \times \int_0^n {}_t p_x dt} = \frac{l_x \times F_T(n \mid X > x)}{l_x \times \int_0^n {}_t p_x dt}
$$

$$
= \frac{l_x \times P(T \leqslant n \mid X > x)}{\int_0^n l_{x+t} dt}
$$

$$
= \frac{{}_n d_x}{{}_n L_x} \tag{3.1.36}
$$

这说明，中心死亡率${}_n m_x$还可以看成是与死亡率${}_n q_x$类似的一种死亡率，只不过与${}_n q_x = \frac{{}_n d_x}{l_x}$相比，其分母产生了变化。下面我们来考查二者之间的关系。

由于

$$
{}_n L_x = \int_0^n l_{x+t} dt = l_x \int_0^n {}_t p_x dt
$$

$$
= l_x [t \cdot {}_t p_x |_0^n - \int_0^n (-t \cdot {}_t p_x \cdot \mu_{x+t}) dt]
$$

$$
= n l_{x+n} + \int_0^n t \cdot l_{x+t} \mu_{x+t} dt \tag{3.1.37}
$$

所以，

$$
\int_0^n t \cdot l_{x+t} \mu_{x+t} dt = {}_n L_x - n l_{x+n} \tag{3.1.38}
$$

式（3.1.38）具有明确的含义，表示在区间$(x, x+n]$上死亡的人在这段时间内生存的总年数，在此基础上，我们定义

$$
{}_n f_x = \frac{{}_n L_x - n l_{x+n}}{n \cdot {}_n d_x} \tag{3.1.39}
$$

则${}_n f_x$表示在这段区间上死亡的人在单位区间上生存的平均年数，这样，我们有

$$
{}_n L_x = n l_{x+n} + n \cdot {}_n d_x \cdot {}_n f_x \tag{3.1.40}
$$

再把$l_{x+n} = l_x - {}_n d_x$代入式（3.1.40），可得：

$$
{}_n L_x = n l_x - n(1 - {}_n f_x) {}_n d_x \tag{3.1.41}
$$

把式（3.1.41）代入式（3.1.36），就可找出${}_n m_x$与${}_n q_x$之间的关系：

$$
{}_n m_x = \frac{{}_n q_x}{n - n(1 - {}_n f_x) {}_n q_x} \tag{3.1.42}
$$

$$ {}_nq_x = \frac{{}_nm_x}{\frac{1}{n} + (1 - {}_nf_x)\,{}_nm_x} \tag{3.1.43} $$

在由人口统计数据编制生命表时，由于资料限制，中心死亡率更容易得到，因此式（3.1.43）在计算死亡率时非常重要。

特殊地，当 $n = 1$ 时，有

$$ m_x = \frac{q_x}{1 - (1 - f_x)q_x} \tag{3.1.44} $$

$$ q_x = \frac{m_x}{1 + (1 - f_x)m_x} \tag{3.1.45} $$

如果再假设死亡时间在整数区间内均匀分布，那么 $f_x = 1/2$（这个结果见例 3－3），式（3.1.45）进一步简化为：

$$ q_x = \frac{2m_x}{2 + m_x} \tag{3.1.46} $$

式（3.1.46）也是编制生命表时常用的一个重要公式。

【例 3－1】 已知 $l_0 = 100\,000$，$l_1 = 97\,408$，$l_5 = 97\,015$，$L_0 = 97\,764$，${}_4L_1 = 388\,713$。求 f_0 和 ${}_4f_1$。

解： $d_0 = l_0 - l_1 = 2\,592$，由式（3.1.39）可知

$$ f_0 = \frac{L_0 - l_1}{d_0} = \frac{97\,764 - 97\,408}{2\,592} = 0.137346 $$

类似地，${}_4d_1 = l_1 - l_5 = 393$，仍由式（3.1.39）可知：

$$ {}_4f_1 = \frac{{}_4L_1 - 4l_5}{4 \cdot {}_4d_1} = \frac{388\,713 - 4 \times 97\,015}{4 \times 393} = 0.415394 $$ ■

【例 3－2】 已知 ${}_5q_{80} = 0.4$，${}_5m_{80} = 0.102$。求 ${}_5f_{80}$。

解： 由式（3.1.43）可解得：

$$ 1 - {}_5f_{80} = \frac{5 \times {}_5m_{80} - {}_5q_{80}}{5 \times {}_5m_{80} \times {}_5q_{80}} = 0.539216 $$

所以，${}_5f_{80} = 0.460784$。 ■

§3.2　相邻整数年龄间的死亡分布

生命表是以整数年龄分组编制的。在保险精算实践中，常常需要相邻整数年间生存或死亡的信息，如 30 岁的人存活半年的概率 ${}_{1/2}p_{30}$，50.5 岁的人 4 个月内死亡的概率 ${}_{1/3}q_{50.5}$，40 岁的人在 40.25 岁的条件概率密度以及任意一个年龄段的中心死亡率，等等。获得这些信息除了利用生命表，还需要对相邻整数年龄间的生存函数作出假设。

常用的几个假设是**死亡时间均匀分布假设**、**死亡力恒定假设**和 **Balducci 假设**，这三种假设分别对应线性插值法、指数插值法和调和插值法。

在3.1.2考虑X岁人剩余寿命的概率密度时，我们曾把初始年龄为x的人的剩余寿命记为$T(x)$，简记为T。在这一节，我们仍然沿用上述的简便记法，把个体的生存时间记为T，在不致混淆的情况下，把相邻的整数年龄区间记为$(x, x+1)$。

3.2.1 死亡时间均匀分布假设

假设死亡时间T在区间$(x, x+1]$内均匀分布，此时生存函数是线性函数，对任意的$0<t<1$，有

$$S(x+t) = (1-t)\cdot S(x) + t\cdot S(x+1) \tag{3.2.1}$$

于是，$${}_tq_x = \frac{S(x)-S(x+t)}{S(x)} = \frac{S(x)-[(1-t)\cdot S(x)+t\cdot S(x+1)]}{S(x)}$$

$$= \frac{t[S(x)-S(x+1)]}{S(x)} = tq_x \tag{3.2.2}$$

$${}_tp_x = 1 - {}_tq_x = 1 - tq_x \tag{3.2.3}$$

$${}_{s|t}q_x = {}_{s+t}q_x - {}_sq_x = tq_x,\quad 0 < s+t \leqslant 1 \tag{3.2.4}$$

式(3.2.4)说明，个体在区间$(x, x+1]$上任何一个长度为t的期间死亡的概率都相同，这也是死亡均匀分布的一个体现。

由式(3.2.4)，有

$${}_tq_{x+s} = \frac{{}_{s|t}q_x}{{}_sp_x} = \frac{tq_x}{1-sq_x},\quad 0\leqslant s\leqslant 1,\quad 0 < s+t \leqslant 1 \tag{3.2.5}$$

另外，$$f_T(t) = -\frac{dS(x+t)/dt}{S(x)} = -\frac{d[(1-t)\cdot S(x)+t\cdot S(x+1)]/dt}{S(x)}$$

$$= \frac{S(x)-S(x+1)}{S(x)} = q_x \tag{3.2.6}$$

又因为$f_T(t) = {}_tp_x\cdot\mu_{x+t}$，所以，

$$\mu_{x+t} = \frac{f_T(t)}{{}_tp_x} = \frac{q_x}{1-tq_x} \tag{3.2.7}$$

μ_{x+t}随着时间递增。

【例3-3】 在死亡时间均匀分布假设下，估计式(3.1.39)定义的f_x。

解：$f_x = \dfrac{L_x - l_{x+1}}{d_x} = \dfrac{l_{x+1}+\frac{1}{2}d_x - l_{x+1}}{d_x} = \dfrac{1}{2}$ ■

3.2.2 死亡力恒定假设

这种情况下，假设生存函数的规律如下：

$$\ln S(x+t) = (1-t)\cdot\ln S(x) + t\cdot\ln S(x+1),\quad 0<t<1 \tag{3.2.8}$$

即 $S(x+t) = S(x)^{1-t} \cdot S(x+1)^t, 0 < t < 1$ (3.2.9)

那么有

$$ {}_tp_x = \frac{S(x+t)}{S(x)} = \frac{S(x)^{1-t} \cdot S(x+1)^t}{S(x)} = \left[\frac{S(x+1)}{S(x)}\right]^t = (p_x)^t $$

即 ${}_tp_x = (p_x)^t$ (3.2.10)

所以，${}_tq_x = 1 - {}_tp_x = 1 - (p_x)^t$ (3.2.11)

而且有

$$ {}_tq_{x+s} = \frac{S(x+s) - S(x+s+t)}{S(x+s)} = \frac{\frac{S(x+s)}{S(x)} - \frac{S(x+s+t)}{S(x)}}{\frac{S(x+s)}{S(x)}} $$

$$ = \frac{(p_x)^s - (p_x)^{s+t}}{(p_x)^s} = 1 - (p_x)^t $$

即 ${}_tq_{x+s} = 1 - (p_x)^t, \quad 0 \leqslant s \leqslant 1, \quad 0 < s+t \leqslant 1$ (3.2.12)

还有

$$ \mu_{x+t} = -\frac{dS(x+t)/dt}{S(x+t)} = -d[\ln S(x+t)]/dt $$

$$ = -d[(1-t) \cdot \ln S(x) + t \cdot \ln S(x+1)]/dt $$

$$ = -\ln\frac{S(x+1)}{S(x)} = -\ln p_x $$

即 $\mu_{x+t} = -\ln p_x$ (3.2.13)

式（3.2.13）说明，个体在区间 $(x, x+1]$ 上的死亡力是常数，所以这种情形称为**死亡力恒定假设**。由于这个常数只与年龄 x 有关，因此不妨记做 μ_x，即 $\mu_x = -\ln p_x$。

另外，$f_T(t) = {}_tp_x \cdot \mu_{x+t} = -(\ln p_x)(p_x)^t$ (3.2.14)

在这一假设下，可以得到如下一些结果：

我们知道，中心死亡率是死亡力的加权平均值，如果死亡力是常数，那么中心死亡率也是常数，这一结论也可通过如下的推导得到验证：

$$ m_x = \frac{d_x}{L_x} = \frac{d_x}{\int_0^1 l_{x+t}dt} = \frac{d_x}{l_x\int_0^1 (p_x)^t dt} = \frac{d_x}{d_x/(-\ln p_x)} = -\ln p_x = \mu_x $$

即 $m_x = \mu_x$ (3.2.15)

由式（3.2.15）可得：

$$ L_x = \frac{d_x}{m_x} = \frac{d_x}{\mu_x} \tag{3.2.16} $$

因此，
$$ \mathring{e}_x = \frac{T_x}{l_x} = \frac{1}{l_x}\sum_{y=x}^{\infty} L_y = \frac{1}{l_x}\sum_{y=x}^{\infty}\frac{d_y}{\mu_y} \tag{3.2.17} $$

【例 3-4】 在死亡力恒定假设下，用 p_x 表示式（3.1.39）定义的 f_x。

解： 由式（3.2.16）和（3.1.39）可得：

$$f_x = \frac{L_x - l_{x+1}}{d_x} = \frac{\dfrac{d_x}{-\ln p_x} - l_{x+1}}{d_x} = -\frac{1}{\ln p_x} - \frac{l_{x+1}}{d_x} = -\frac{1}{\ln p_x} - \frac{p_x}{1-p_x}$$ ■

3.2.3 Balducci 假设

这种情况下，假设生存函数的规律如下：

$$\frac{1}{S(x+t)} = \frac{1-t}{S(x)} + \frac{t}{S(x+1)}, 0 < t < 1 \tag{3.2.18}$$

由式（3.2.18）可得：

$$S(x+t) = \left[\frac{1}{S(x)} + t\left(\frac{1}{S(x+1)} - \frac{1}{S(x)}\right)\right]^{-1} \tag{3.2.19}$$

因此，
$$\frac{1}{{}_tp_x} = \frac{S(x)}{S(x+t)} = S(x)\left[\frac{1}{S(x)} + t\left(\frac{1}{S(x+1)} - \frac{1}{S(x)}\right)\right]$$

$$= 1 + t\left(\frac{1}{p_x} - 1\right) \tag{3.2.20}$$

则
$${}_tp_x = \frac{1-q_x}{1-(1-t)q_x} \tag{3.2.21}$$

$${}_tq_x = \frac{tq_x}{1-(1-t)q_x} \tag{3.2.22}$$

而且有
$${}_tp_{x+s} = \frac{p_x + sq_x}{p_x + (s+t)q_x}, 0 \leqslant s \leqslant 1, 0 < s+t \leqslant 1 \tag{3.2.23}$$

$${}_tq_{x+s} = \frac{tq_x}{p_x + (s+t)q_x}, 0 \leqslant s \leqslant 1, 0 < s+t \leqslant 1 \tag{3.2.24}$$

其中，式（3.2.22）的推导留做练习。

另外，由 $\mu_{x+t} = -\dfrac{d({}_tp_x)/dt}{{}_tp_x}$ 和式（3.2.22），有

$$\mu_{x+t} = \frac{q_x}{1-(1-t)q_x} = \frac{q_x}{p_x + tq_x} \tag{3.2.25}$$

式（3.2.25）说明 μ_{x+t} 随着时间递减。事实上，在非整数年龄内死亡力递减的情况下，式（3.2.25）对于估计死亡概率非常重要。

另外，还有

$$f_T(t) = {}_tp_x \cdot \mu_{x+t} = \frac{q_x(1-q_x)}{[1-(1-t)q_x]^2} \tag{3.2.26}$$

特殊地，当 $t = 1 - s$ 时，有

$${}_{1-s}q_{x+s} = (1-s)q_x, 0 \leqslant s \leqslant 1 \tag{3.2.27}$$

意大利精算学家 Gaetano Balducci 在他的许多论文中使用这一假设，因此称这一假设为“Balducci 假设”。

【例 3-5】 假设相邻整数年龄间的死亡服从 Balducci 分布，求 $\mathring{e}_x$。

解： 由式（3.2.27），在 Balducci 假设下，有 ${}_{1-s}p_{x+s} = p_x + sq_x$，$0 \leqslant s \leqslant 1$。

那么，

$$L_x = \int_0^1 l_{x+s} ds = l_{x+1} \int_0^1 \frac{l_{x+s}}{l_{x+1}} ds$$

$$= l_{x+1} \int_0^1 \frac{1}{{}_{1-s}p_{x+s}} ds = l_{x+1} \int_0^1 \frac{1}{p_x + sq_x} ds = -\frac{l_{x+1} \ln p_x}{q_x}$$

$$\mathring{e}_x = \frac{T_x}{l_x} = \frac{1}{l_x}\sum_{y=x}^{\infty} L_y = -\frac{1}{l_x}\sum_{y=x}^{\infty}\frac{l_{y+1}\ln p_y}{q_y} = -\sum_{y=x}^{\infty}\frac{{}_{y-x+1}p_x \cdot \ln p_y}{q_y}$$ ■

【例 3－6】 在某生命表中，$l_x = 1\,000$，$l_{x+1} = 900$，分别在死亡时间均匀分布假设、死亡力恒定假设、Balducci 假设下估计 m_x。

解： 由已知 $d_x = 100$，$\ln p_x = -0.10536$。

在死亡时间均匀分布假设下：$L_x = l_x - \frac{1}{2}d_x = 950$，则 $m_x = \frac{d_x}{L_x} = 0.10526$；

在死亡力恒定假设下：$m_x = \mu_x = -\ln p_x = 0.10536$；

在 Balducci 假设下：$L_x = -\frac{l_{x+1}\ln p_x}{q_x} = 948.24$，因此 $m_x = 0.10546$。 ■

表 3－2 是对上述三种假设下估计结果的总结。

表 3－2　　不同假设下的生命表函数及其关系式

函数	线性（死亡时间均匀分布假设）	指数（死亡力恒定假设）	调和（Balducci 假设）
$S(x+t)$	$S(x+t) = (1-t)S(x) + tS(x+1)$	$S(x+t) = [S(x)]^{1-t} \cdot [S(x+1)]^t$	$\frac{1}{S(x+t)} = \frac{1-t}{S(x)} + \frac{t}{S(x+1)}$
${}_tq_x$	tq_x	$1-(p_x)^t$	$\frac{tq_x}{p_x+tq_x}$
${}_tp_x$	$1-tq_x$	$(p_x)^t$	$\frac{p_x}{p_x+tq_x}$
${}_tq_{x+s}$	$\frac{tq_x}{1-sq_x}$	$1-(p_x)^t$	$\frac{tq_x}{p_x+(t+s)q_x}$
${}_{1-t}q_{x+t}$	$\frac{(1-t)q_x}{1-tq_x}$	$1-(p_x)^{1-t}$	$(1-t)q_x$
μ_{x+t}	$\frac{q_x}{1-tq_x}$	$-\ln p_x$	$\frac{q_x}{p_x+tq_x}$
$f_T(t)$	q_x	$(-\ln p_x)\cdot(p_x)^t$	$\frac{p_xq_x}{(p_x+tq_x)^2}$
L_x	$l_x - \frac{1}{2}d_x$	$-\frac{d_x}{\ln p_x}$	$-\frac{l_{x+1}\ln p_x}{q_x}$
m_x	$\frac{q_x}{1-\frac{1}{2}q_x}$	$-\ln p_x$	$\frac{(q_x)^2}{-p_x\cdot\ln p_x}$

§3.3 选择—终极生命表

3.3.1 经验生命表的种类

生命表有很多种类，其用途各不相同。国民生命表根据全体国民的死亡数据编制而成，可以概括描述国民总体的寿命状况；经验生命表由人寿保险公司根据投保人的死亡记录编制，通常由寿险公司专用。

经验生命表按照经验统计资料性质的不同，可以分为选择生命表、终极生命表和综合生命表。

1. 选择生命表。成为寿险保单的被保险人通常要经过体检，他们的死亡率在最初几年内较一般水平低，这最初的几年被保险公司定为选择期。**仅根据处于选择期的被保险人死亡数据编制的生命表称为“选择生命表”。**选择生命表可以显示保险公司通过体检等手段选择被保险人的效力，一般来说，被保险人的投保年龄和投保时间成为影响死亡率的最主要因素，除此之外，性别、是否吸烟等因素也会明显影响死亡率。

在选择生命表中也可以列出各年龄的生存人数 $l_{[x]}$，也有与综合生命表中类似的关系式。如

$$l_{[x-k]+r+k}=l_{[x]+r}=l_{x+r}\ ,\ d_{[x]+n}=l_{[x]+n}-l_{[x]+n+1}$$

$$q_{[x]+n}=\frac{d_{[x]+n}}{l_{[x]+n}}\ ,\ p_{[x]+n}=\frac{l_{[x]+n+1}}{l_{[x]+n}}$$

$${}_{m|}q_{[x]+n}=\frac{l_{[x]+n+m}-l_{[x]+n+m+1}}{l_{[x]+n}}$$

2. 终极生命表。**根据已经渡过选择期的被保险人的死亡数据编制的生命表称为“终极生命表”。**

3. 综合生命表。**不考虑选择期，由全体被保险人的死亡数据编制的生命表称为“综合生命表”。**

3.3.2 选择—终极生命表

由于保险公司核保的选择，一组被保险人的死亡率不仅随投保年龄变动，而且随投保时间变动。以 $q_{[x]+n}$ 表示 x 岁投保、经过 n 年后在 $(x+n)$ 岁死亡的概率，有

$$q_{[x]}<q_{[x-1]+1}<q_{[x-2]+2}<\cdots \tag{3.3.1}$$

通常选择效力会随时间增加逐渐消失，即投保时间 n 越长，$q_{[x-n]+n}$ 与 $q_{[x-n+1]+n-1}$ 越趋于相等，选择期最长也不会超过 15 年，我们把选择期记为 r 年。

对于任何一个特定年龄的被保险人来说，如果保险期间开始时他的年

龄为整数 x，那么在接下来的每一个整数年内，他的死亡率为 $q_{[x]+n}$，$n=0$，1，2，…，一般地，有

$$q_{[x]+n} < q_{x+n}, \qquad n = 0, 1, 2, \cdots, r-1 \tag{3.3.2}$$

过了 r 年的选择期，选择的效力消失，即

$$q_{[x]+n} = q_{x+n}, \qquad n = r, r+1, \cdots \tag{3.3.3}$$

根据 $q_{[x]+n}$，$n=0$，1，2，…，$r-1$ 编制的生命表就是选择生命表，由选择效力消失后的死亡率编制的生命表就是终极生命表，习惯上将终极表并列在选择表的右边，这样的生命表就是选择—终极生命表，如表 3－3，为简明起见，这里假定选择期为 5 年。

表 3－3　　　　选择—终极生命表例表

$[x]$	选择表					终极表	
	$q_{[x]}$	$q_{[x]+1}$	$q_{[x]+2}$	$q_{[x]+3}$	$q_{[x]+4}$	$q_{[x]+5}$	$x+5$
…	…	…	…	…	…	…	…
[70]	0.0175	0.0249	0.0313	0.0388	0.0474	0.0545	75
[71]	0.0191	0.0272	0.0342	0.0424	0.0518	0.0596	76
[72]	0.0209	0.0297	0.0374	0.0463	0.0566	0.0652	77
[73]	0.0228	0.0324	0.0409	0.0507	0.0620	0.0714	78
[74]	0.0249	0.0354	0.0447	0.0554	0.0678	0.0781	79
[75]	0.0273	0.0387	0.0489	0.0607	0.0742	0.0855	80
[76]	0.0298	0.0424	0.0535	0.0664	0.0812	0.0936	81
[77]	0.0326	0.0464	0.0586	0.0727	0.0889	0.1024	82
[78]	0.0357	0.0508	0.0641	0.0796	0.0973	0.1121	83
[79]	0.0391	0.0556	0.0702	0.0871	0.01065	0.1227	84
…	…	…	…	…	…	…	…

在表 3－3 中，$q_{[72]+3} = 0.0463$，表示投保年龄为 72 岁的被保险人投保 3 年后的死亡率是 0.0374，但是，过了选择期后，投保年龄为 72、73、74 和 75 岁的人在 80 岁的死亡率都是 0.0855。

表 3－4　　选择生命表例表

$[x]$	$l_{[x]}$	$l_{[x]+1}$	$l_{[x]+2}$	$l_{[x]+3}$	$l_{[x]+4}$	$x+4$
35	1 000	998	994	990	985	39
36	996	993	990	987	983	40
37	995	991	987	982	977	41
38	992	987	981	975	968	42
39	988	983	975	967	958	43
40	982	977	972	963	951	44

从选择表的左下至右上可以引出一些对角线，同一条线上的数据表示在不同年龄投保但处在同一到达年龄的被保险人的死亡率，以表 3－3 中划线数据为例，在到达年龄为 74 岁这条线上可以查到如下的死亡率：$q_{[74]} = 0.0249$，$q_{[73]+1}$

$=0.0324$，$q_{[72]+2}=0.0374$，$q_{[71]+3}=0.0424$，$q_{[70]+4}=0.0474$。

【例 3－7】 根据如表 3－4 所示的选择生命表，计算$_3p_{[35]+1}$、$_3p_{[38]}$、$_2q_{[37]+2}$、$_{2|}q_{[36]}$。

解：$_3p_{[35]+1}=\dfrac{l_{[35]+4}}{l_{[35]+1}}=\dfrac{l_{39}}{l_{[35]+1}}=\dfrac{985}{998}=0.986974$

$$_3p_{[38]}=\frac{l_{[38]+3}}{l_{[38]}}=\frac{l_{[38]+3}}{l_{[38]}}=\frac{975}{992}=0.982863$$

$$_2q_{[37]+2}=\frac{l_{[37]+2}-l_{[37]+4}}{l_{[37]+2}}=\frac{l_{[37]+2}-l_{41}}{l_{[37]+2}}=\frac{987-977}{987}=0.010132$$

$$_{2|}q_{[36]}=\frac{l_{[36]+2}-l_{[36]+3}}{l_{[36]}}=\frac{990-987}{996}=0.003012$$

■

【例 3－8】 在选择期为 1 年的某生命表中，下列关系对于所有年龄都成立：$_{0.5}q_{[x]}=0.33q_x$，$_{0.5}q_{[x]+0.5}=0.5q_x$，试用 p_x 表示 $p_{[x]}$。

解：

$$\begin{aligned}p_{[x]}&=({}_{0.5}p_{[x]})({}_{0.5}p_{[x]+0.5})=(1-{}_{0.5}q_{[x]})(1-{}_{0.5}q_{[x]+0.5})\\&=(1-0.33q_x)(1-0.5q_x)=[1-0.33(1-p_x)][1-0.5(1-p_x)]\\&=(0.67+0.33p_x)(0.5+0.5p_x)=0.335+0.5p_x+0.165p_x^2\end{aligned}$$

■

习　题

1. 某生命表生存概率 p_x 的值如表 3－5 所示。

表 3－5

X	0	1	2	3	4
p_x	0.9	0.8	0.6	0.3	0.0

（1）计算对应于 $x=0$，1，2，3，4，5 的 $S(x)$的值；

（2）以 10 000 为基数，推导出表示 d_x 和 l_x 值的生命表；

（3）该表中的 ω 是多少？

2. 由上题中的生命表计算下列各值：（1）$_3d_0$；（2）$_2q_1$；（3）$_3p_1$；（4）$_3q_2$。

3. 已知新生儿在 x 岁与（$x+1$）岁之间死亡的无条件概率为$_{x|}q_0$，要求：

（1）分别用 $S(x)$与 l_x 定义$_{x|}q_0$；

（2）证明 $\sum_{x=0}^{\omega-1}{}_{x|}q_0=1$。

4. 某生命表由生存函数 $S(x)=\dfrac{c-x}{c+x}$，$0\leqslant x\leqslant c$ 构造，令 $l_0=100\,000$，且已知 $l_{35}=44\,000$。（1）求这个生命表中的 ω；（2）求新生儿活到 60 岁的

概率；（3）求 10 岁的人在 30 岁与 45 岁之间死亡的概率。

5. 如果 $\mu_x = \dfrac{2}{x+1} + \dfrac{2}{100-x}$, $0 \leqslant x < 100$。求 $l_0 = 10\ 000$ 时，在该生命表中 1 ~ 4 岁之间的死亡人数。

6. 求 $\dfrac{\partial({}_tp_x)}{\partial x}$。

7. 如果 $l_x = 2\ 500 \times (64 - 0.8x)^{\frac{1}{3}}$, $0 \leqslant x \leqslant 80$，求 ${}_{10}m_{70}$ 的值。

8. 证明：$e_x = p_x(1 + e_{x+1})$。

9. 给定某一生命表中的下列数据：$l_{50} = 80\ 000$，$l_{74} = 42\ 693$，$l_{75} = 40\ 280$，$l_{70} = 37\ 480$，在下列假设下求一个 50 岁人未来寿命的中位数：（1）死亡时间均匀分布；（2）死亡力恒定。

10. 对于选择年龄为 0 岁的 3 年选择期的选择—终极表，在下列条件下求 $l_{[0]}$：$l_6 = 90\ 000$, $q_{[0]} = \dfrac{1}{6}$, ${}_5p_{[1]} = \dfrac{4}{5}$, $d_x = 5\ 000(x \geqslant 3)$, ${}_3p_{[0]+1} = \dfrac{9}{10} \cdot {}_3p_{[1]}$。

11. 已知 $l_{30} = 98\ 617$，$l_{40} = 97\ 952$。在以下两个假设下计算 ${}_5q_{30}$：

（1）在 30 ~ 40 岁之间死亡时间均匀分布；

（2）在 30 ~ 40 岁之间死亡力恒定。

（3）分别在（1）、（2）两个假设下计算 10 万个新生儿中活到 35 岁的人数。

12. 假设整数年 y 和 $y+1$ 之间的死亡力 μ_y 是恒定的，令 T_x 是 x 岁以后的生存年，$S_x(t)$ 是 T_x 的生存函数。证明：$\mu_y = \ln[S_x(y-x)] - \ln[S_x(y+1-x)]$。

13. 已知一组动物的死亡率为 $q_x = 0.1$，分别在三种假设下计算 m_x：（1）死亡时间均匀分布假设；（2）死亡力恒定假设；（3）Balducci 假设。

14. 已知 $l_x = 1\ 000\left(1 - \dfrac{x}{120}\right)$，计算下面各项的值：

（1）l_0，l_{120}，d_{33}，${}_{30}q_{20}$，${}_{20}p_{30}$；

（2）25 岁的人至少存活 20 年，最多存活 25 年的概率；

（3）3 个 25 岁的人均存活到 80 岁的概率。

15. 已知 ${}_{1|}q_{x+1} = 0.095$，${}_{2|}q_{x+1} = 0.171$，$q_{x+3} = 0.2$，计算 $q_{x+1} + q_{x+2}$。

16. 已知 $l_{65} = 100$，$l_{66} = 80$。试分别在死亡时间均匀分布、死亡力恒定和 Balducci 假设下计算 $l_{65.5}$。

17. 表 3 – 6 是一个选择期为一年的生命表。假设每一年死亡均匀分布，计算 $\mathring{e}_{[81]}$。

18. 已知：$\mu(x) = \begin{cases} 0.05, & 50 \leqslant x < 60 \\ 0.04, & 60 \leqslant x < 70 \end{cases}$，计算 ${}_{4|14}q_{50}$。

19. 某两年期的疾病治疗项目信息如下：（1）只有10%的人生存至第二年年底；（2）每个整数年的死亡力恒定；（3）第二年的死亡力是第一年的3倍。计算一个在第三个月末还存活的项目参加者在第21个月末死亡的概率。

表3-6

x	$l_{[x]}$	$d_{[x]}$	l_{x+1}	$\mathring{e}_{[x]}$
80	1 000	90	—	8.5
81	920	90	—	—

20. 证明：$L_{x+1} = L_x \cdot e^{-\int_x^{x+1} m_y dy}$。

第四章　理赔额和理赔次数的分布

学习目标

- □ 了解以下概念：理赔额、损失额、免赔额、保单限额和比例赔偿
- □ 了解不同的赔偿方式对理赔额和理赔次数造成的影响
- □ 熟悉常见的损失额分布以及不同赔偿方式下理赔额的分布
- □ 熟悉单个保单理赔次数的分布以及$(a,b,0)$分布类和$(a,b,1)$分布类
- □ 熟悉不同结构函数下保单组合理赔次数的分布以及相关性保单组合理赔次数的分布
- □ 掌握并运用各种条件下理赔额与理赔次数的分布解决实际问题

§4.1　损失额分布

4.1.1　理赔额与损失额

在非寿险业务的经营中，保单产生理赔额大致包括两个步骤：（1）保险事故发生，造成财产损失或人身伤亡；（2）被保险人提出索赔，保险公司进行理赔。但是，并不是所有的保险事故必然引起索赔，而且保险公司的理赔额也并不总是等于实际的损失额。

损失额和理赔额是两个不同但又密切相关的概念，**损失额**是指承保标的发生实际损失金额的大小，而**理赔额**是指保险公司按照保单条款所实际支付的金额，也可称为“赔付额”。理赔额通常小于实际损失额。一般来说，理赔分为两类：完全理赔和部分理赔。在完全理赔中，理赔额就是保险事故的实际损失额；在部分理赔中，理赔额可能会低于实际损失额。部分理赔涉及的基本概念如下：

1. 免赔额（Deductible）。是指保单规定的最低起赔额，当损失额低于这一额度时，保险公司不赔偿，保险公司只赔偿高出的部分。

2. 保单限额（Policy Limit）。是指保单约定的最高赔偿金额。当损失金额超过保单限额时，被保险人将只获得最高赔偿额，超出部分由被保险人承担。

保险公司设置免赔额和保单限额的目的，是建立一种与被保险人共担风险的机制，从而有效控制保险赔款支出。除此之外，免赔额还可以提高

被保险人的安全意识，减少理赔次数并降低保险经营的费用。

如果保单同时规定了免赔额 d 和保单限额 D，则被保险人实际所能得到的最高赔偿金额为 $D-d$。

3. 比例赔付。是指在保单中约定一个比例常数 k, $0<k<1$，当保险事故的实际损失额为 X 时，保险公司支付赔偿金 kX，剩余的损失额 $(1-k)X$ 由被保险人自己承担。当保单中同时设有免赔额 d、保单限额 D 和赔付比例 k 时，被保险人最终得到的理赔额最高为 $k(D-d)$。

对保险公司来说，理赔额比损失额更值得关心，因为费率厘定、准备金计提及再保险安排等精算问题都以理赔额的分布为依据。但是，理赔额是在损失额的基础上定义的，只有对损失额分布有充分的了解，才能获得理赔额的分布。因此，我们有必要先研究损失额的分布，再研究理赔额的分布。

4.1.2 常见的损失额分布

由于损失额和理赔额的不确定性，因此常用随机变量来描述它们，如果再考虑时间因素，就可以用随机过程来衡量。对于一个随机变量来说，把握其特征规律最重要的就是它的分布。除了分布之外，了解其数字特征或各阶矩，对于把握随机变量的特征也有重要意义。

我们首先考虑单个保单或个体保单损失额的分布。从直观上讲，单个保单的损失额应该具有下面的分布特征：

(1) 损失额是非负的，因此 $P(X\geqslant 0)=1$；

(2) 损失额应该是连续变化的，因此 $f(x)$ 是连续的；

(3) 损失额较小的保险事故发生的可能性较大，而损失额较大的保险事故发生的可能性较小，但不可以忽略。直观看来，损失额概率密度函数的尾部较厚，如图 4-1 所示。

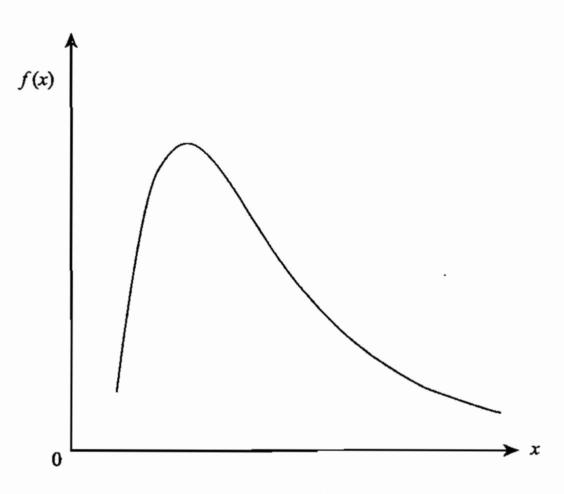


图 4-1 单个保单损失额的概率密度

满足上述性质的随机变量很多，常见的分布有指数分布、伽玛分布、对数正态分布、帕累托分布和韦伯分布，我们下面会对这些分布作简单介绍。

在介绍分布之前，由于矩母函数对于求解随机变量各阶矩的重要性，我们首先介绍矩母函数。

定义 4－1　随机变量 X 的分布函数为 $F(x)$，其矩母函数为：

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{+\infty} e^{tx} dF(x),\ t \geq 0$$

由定义 4－1 可以看出，矩母函数在原点总是有定义的，且 $M_X(0) = 1$。如果 X 的矩母函数在原点的某邻域 $|t| < r(r > 0)$ 内存在，则在此邻域内，$M_X(t)$ 具有如下性质：

性质 1　在 $|t| < r$ 内，X 的分布函数由矩母函数 $M_X(t)$ 唯一确定；比如有两个分布函数 $F_1(x)$ 和 $F_2(x)$，若它们对应的矩母函数相同，则有 $F_1(x) \equiv F_2(x)$。

性质 2　记 X 的 k $(k = 1, 2, \cdots)$ 阶原点矩为 $p_k = E(X^k)$，则有

$$p_k = M_X^{(k)}(0), \qquad k = 1, 2, \cdots$$

并且矩母函数 $M_X(t)$ 还可以进行如下的 Taylor 展开：

$$M_X(t) = \sum_{k=0}^{\infty} p_k \frac{t^k}{k!},\ |t| < r$$

性质 3　若 $X_1, X_2, \cdots, X_n$ 为相互独立的随机变量，则这些随机变量的和

$$S = X_1 + X_2 + \cdots + X_n$$

的矩母函数为各随机变量矩母函数的乘积：

$$M_S(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t)$$

性质 4　若 $Y = aX + b$，a，b 为常数，则随机变量 Y 的矩母函数为：

$$M_Y(t) = e^{bt} M_X(at)$$

性质 1 可看做是应用矩母函数处理分布函数的唯一性定理；性质 2 至性质 4 则是关于矩母函数最常用的性质，这四条性质的数学证明可参阅有关的概率论教材。

由于我们经常要用到正态分布，因此把正态分布也列在这里：

1. 正态分布。若随机变量 X 的密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ -\infty < x < +\infty,\ -\infty < \mu < +\infty,\ \sigma > 0$$

称 X 服从参数为 (μ, σ^2) 的正态分布。

这里我们仅以正态分布为例给出矩母函数的求解过程，其余分布相应的求解过程从略，请读者自己证明。正态分布的矩母函数为：

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{+\infty} e^{tx} f(x) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{tx} \times e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{\mu t + \frac{1}{2}\sigma^2 t^2} \times e^{-\frac{[x-(\mu+\sigma^2 t)]^2}{2\sigma^2}} dx = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

标准正态随机变量 $X \sim N(0, 1)$ 的矩母函数为：

$$M_X(t) = e^{\frac{1}{2}t^2}$$

2. 指数分布。若随机变量 X 的密度函数为

$$f(x) = \frac{1}{\theta}e^{-\frac{1}{\theta}x},\ x > 0,\ \theta > 0$$

其中 θ 是常数，则称 X 服从参数为 θ 的指数分布。指数分布的矩母函数为：

$$M_X(t) = (1 - \theta t)^{-1},\ 0 < t < \frac{1}{\theta}$$

指数分布的期望、方差和 k 阶原点矩为：

$$E(X) = \theta,\ Var(X) = \theta^2,\ E(X^k) = \Gamma(k+1)\theta^k$$

我们在第二章中曾经介绍过指数分布，它可以用于拟合个体寿命。

3. 伽玛分布。若随机变量 X 的密度函数为

$$f(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\Gamma(\alpha)\theta^{\alpha}},\ \alpha > 0,\ \theta > 0,\ x > 0$$

其中 $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}e^{-t}dt$，对正整数 k，$\Gamma(k+1) = k!$，则称 X 服从参数为 (α,θ) 的伽玛分布。当 $\alpha = 1$ 时，伽玛分布退化为指数分布。当参数 $\theta = 1$ 时，伽玛分布称为标准伽玛分布，其分布函数记为：

$$\Gamma(x;\alpha) = \frac{1}{\Gamma(\alpha)}\int_0^{x} t^{\alpha-1}e^{-t}dt$$

伽玛分布的矩母函数为：

$$M_X(t) = E(e^{tX}) = (1 - \theta t)^{-\alpha},\qquad 0 < t < \frac{1}{\theta}$$

利用矩母函数的性质，容易计算得伽玛分布的期望、方差和 k 阶原点矩分别为：

$$E(X) = \alpha\theta$$

$$Var(X) = \alpha\theta^2$$

$$E(X^k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)}\theta^k,\qquad k > -\alpha$$

从期望和方差的表达式可以看出，当 α 固定时，θ 越大，伽玛分布的期望值和方差也越大，密度函数将向右偏移，同时趋于平缓（如图 4－2 所示）。而 θ 固定时，α 越大则期望与方差越大，密度函数也向右偏移（如图 4－3 所示）。

伽玛随机变量一个常用的性质是**可加性**。如果 X_1，…，X_n 是分别服从参数为 $(\alpha_i,\ \theta)$ 的互相独立的伽玛随机变量，则 $Y = \sum_{i=1}^{n} X_i$ 服从参数为 $(\sum_{i=1}^{n}\alpha_i,\ \theta)$ 的伽玛分布。特别地，若 X_1，…，X_n 是独立且参数都为 θ 的指数随机变量，则 $Y = \sum_{i=1}^{n} X_i$ 服从参数为 (n,θ) 的伽玛分布。

4. 帕累托分布。若随机变量 X 的概率密度函数为

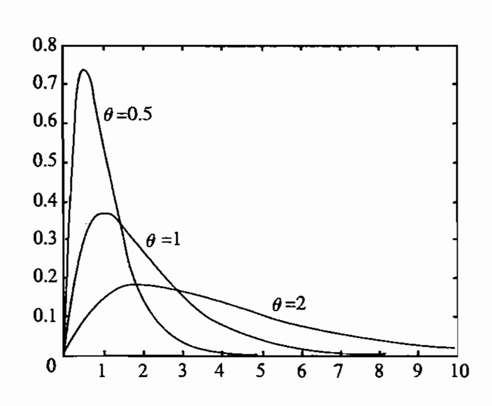


图4-2 伽玛随机变量的分布密度（$\alpha=2$）

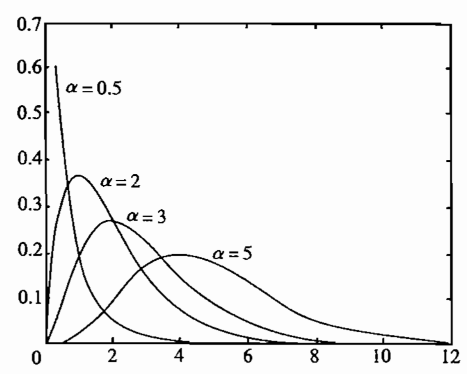


图4-3 伽玛随机变量的分布密度（$\theta=2$）

$$f(x) = \frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}}, \qquad x > 0, \alpha > 0, \theta > 0$$

则称 X 服从参数为（α,θ）的帕累托分布。帕累托分布的分布函数为 $F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha}$。帕累托分布的矩母函数无简单表达式。

利用数学归纳法，可以求得帕累托分布的 k 阶原点矩为：

$$E(X^k) = \frac{\theta^k\Gamma(k+1)\Gamma(\alpha-k)}{\Gamma(\alpha)}, -1 < k < \alpha$$

于是， $E(X) = \frac{\theta}{\alpha-1}, \alpha > 1$；$Var(X) = \frac{\theta^2\alpha}{(\alpha-1)^2(\alpha-2)}, \alpha > 2$

观察帕累托分布的期望与方差表达式可以看出，

$$Var(X) = [E(X)]^2\frac{\alpha}{\alpha-2}, \alpha > 2$$

当 $\alpha\to\infty$ 时，$Var(X)/[E(X)]^2$ 趋于1。事实上，可以证明，当均值 $\mu = E(X)$ 保持不变，令 $\alpha\to\infty$，则该帕累托分布收敛到指数分布。我们把这一结论留做习题。

5. 对数正态分布。假设随机变量 X 取对数后服从正态分布 $N(\mu, \sigma^2)$，则称 X 服从参数为（μ, σ^2）的对数正态分布，记作 $X \sim LN(\mu, \sigma^2)$。对数正态分布的密度函数为：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \qquad x > 0$$

分布函数为 $F(x) = \Phi\left[\frac{(\ln x - \mu)}{\sigma}\right]$，为标准正态随机变量 $N(0, 1)$ 的分布函数。对数正态分布的矩母函数无简单表达式，但是我们可以利用正态分布的矩母函数，计算对数正态的 k 阶原点矩为：

$$E(X^k) = e^{\mu k+\frac{\sigma^2}{2}k^2}$$

于是，$E(X) = e^{\mu+\frac{\sigma^2}{2}}$，$Var(X) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$。可以看出，对数正态分布的期望和方差都是参数$\mu$和$\sigma^2$的增函数。对给定$\mu$，当$\sigma^2 \to 0$时，$E(X) \to \exp(\mu)$，$Var(X) \to 0$。

对数正态分布还有一个非常好的性质，即可乘性。设a和b为正实数，X服从参数为μ和σ^2的对数正态分布，则$Y = aX^b$仍服从对数正态分布，其参数为（$b\mu + \ln a$，$b^2\sigma^2$）。

6. 韦伯分布。若随机变量X的概率密度函数为

$$f(x) = \frac{\gamma}{\theta}x^{(\gamma-1)}e^{-\frac{1}{\theta}x^\gamma}, \qquad x > 0, \theta > 0, \gamma > 0$$

则称X服从参数为（γ，θ）的韦伯分布。当$\gamma = 1$时，韦伯分布为指数分布。韦伯分布的矩母函数也无简单表达式。韦伯分布的k阶原点矩$E(X^k) = \Gamma(1 + k/\gamma)\theta^{k/\gamma}$，$k > -\gamma$，于是期望和方差为：

$$E(X) = \Gamma(1 + 1/\gamma)\theta^{1/\gamma},$$

$$Var(X) = \Gamma(1 + 2/\gamma)\theta^{2/\gamma} - [\Gamma(1 + 1/\gamma)\theta^{1/\gamma}]^2$$

§4.2 理赔额分布

4.2.1 带有免赔额的理赔额分布

记X为保险事故造成的损失额，其分布函数为$F_X(x)$，记Y为保险公司根据条款约定对保单支付的理赔额，其分布函数为$F_Y(y)$，赔款Y可以看做对实际损失额随机变量X的一种修正，类似于我们在第二章中对剩余寿命X作的各种截断。

在考察理赔额的分布中，有两种常用的修正形式：一种是“左截断”，也就是第二章中所说的“截下尾”，即考虑免赔额对理赔额的影响。假设保单规定了免赔额d，理赔额Y可以看做对损失额X的截断：

$$Y = (X - d) \mid (X > d)$$

Y表示在$X > d$的条件下，随机变量$X - d$的分布，Y是一个条件随机变量，其取值范围是$y > 0$。

下面我们来求Y的分布。记Y的分布函数为$F_Y(y)$，当$y > 0$时，

$$F_Y(y) = P(Y \leqslant y) = P(X - d \leqslant y \mid X > d)$$

$$= \frac{P(d < X < y + d)}{P(X > d)} = \frac{F(y + d) - F(d)}{1 - F(d)}$$

因此，Y的概率密度函数为：

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{f(y + d)}{1 - F(d)}, y > 0$$

4.2.2　带有保单限额的理赔额分布

另一种对损失额的修正形式为“右删失”。若规定了保单限额为 l，则 Y 为理赔额，这时有

$$Y=\begin{cases}X, & X\leqslant l\\ l, & X>l\end{cases}$$

Y 是连续随机变量与离散随机变量的混合，容易计算出 Y 的分布密度函数为：

$$f_Y(y)=f(y),\ y<l$$

在 l 点的概率为 $P(Y=l)=1-F(l)$。

定义 4－2　设 X 是一个随机变量，给定实数 l，定义有限期望函数

$$E(X\wedge l)=\int_{-\infty}^{l}xf(x)dx+l[1-F(l)]$$

其中 $F(x)$ 和 $f(x)$ 为 X 的分布函数和密度函数，$X\wedge l$ 的含义是 X 和 l 的最小值。

对于非负随机变量 X，$E(X\wedge d)$ 对任意 $d>0$ 都存在，

$$E(X\wedge d)=\int_{0}^{d}xf(x)dx+d[1-F(d)]=\int_{0}^{d}[1-F(y)]dy$$

尽管上面第二个等式的证明不难，但是我们经常会用到这个公式。显然，当 d 趋于无穷时，$E(X\wedge d)=E(X)$。

此外，在保单限额为 l 时，理赔额的期望 $E(Y)=E(X\wedge l)$。

定义 4－3　设 X 是一个随机变量，给定实数 d，定义剩余期望函数

$$e_X(d)=\int_{d}^{+\infty}\frac{(x-d)f(x)}{1-F(d)}dx=\frac{E(X)-E(X\wedge d)}{1-F(d)}\qquad(4.2.1)$$

其中 $F(x)$ 和 $f(x)$ 分别为 X 的分布函数和密度函数。

$e_X(d)$ 表示随机变量 X 比 d 高出的平均水平，即在 $X>d$ 的条件下，$X-d$ 的期望值，即免赔额为 d 时理赔额的期望。如果 X 表示产品的使用寿命，$e_X(d)$ 表示产品使用 d 的时间长度后剩余的平均寿命，相当于第二章中的 $\mathring{e}_d$，因此 $e_X(d)$ 反映了随机变量的尾部性质，是对随机变量做了“掐头”的处理；而 $E(X\wedge d)$ 则反映了随机变量的前部性质，是作“去尾”的处理。

【例 4－1】　设某险种的损失额 X（万元）具有密度函数 $f(x)=\dfrac{324}{(3+x)^5}$，$x>0$，假定免赔额 d 为 0.5 万元，求理赔额 Y 的期望。

解： 首先理赔额 Y 可以表示为：

$$Y=(X-0.5)\mid(X>0.5)$$

那么，由（4.2.1）式可知：

$$E(Y)=E(X-0.5\mid X>0.5)=\frac{E(X)-E(X\wedge 0.5)}{1-F_X(0.5)}$$

而由已知

$$F_X(x)=\int_0^x \frac{324}{(3+y)^5}dy=-\frac{324}{4}(3+y)^{-4}\Big|_0^x=1-\frac{81}{(3+x)^4}$$

所以，$1-F_X(0.5)=81/(3.5)^4=0.5398$，又 $E(X)=1$，

$$E(X\wedge 0.5)=\int_0^{0.5}[1-F_X(y)]dy=1-\frac{27}{(3+0.5)^3}=0.3703$$

所以，$E(Y)=\dfrac{E(X)-E(X\wedge 0.5)}{1-F_X(0.5)}=\dfrac{1-0.3703}{0.5398}=1.1665$ ■

【例 4-2】 假设实际损失额 X 服从参数 $\alpha=2$ 和参数 θ 的帕累托分布，已知 $3e_X(100)=5e_X(50)$，求 $e_X(150)$。

解： 对于帕累托分布，其剩余期望函数

$$e_X(d)=\frac{E(X)-E(X\wedge d)}{1-F(d)}$$

$$=\frac{\dfrac{\theta}{\alpha-1}-\dfrac{\theta}{\alpha-1}\left[1-\left(\dfrac{\theta}{d+\theta}\right)^{\alpha-1}\right]}{1-\left[1-\left(\dfrac{\theta}{d+\theta}\right)^{\alpha}\right]}=\frac{\dfrac{\theta}{\alpha-1}\left(\dfrac{\theta}{d+\theta}\right)^{\alpha-1}}{\left(\dfrac{\theta}{d+\theta}\right)^{\alpha}}=\frac{d+\theta}{\alpha-1}$$

又由已知 $\alpha=2$，因此 $e_X(d)=d+\theta$，所以 $3(100+\theta)=5(50+\theta)$，从而得 $\theta=25$，因此 $e_X(d)=150+\theta=175$。 ■

4.2.3 带有免赔额、保单限额和比例赔偿的理赔额分布

定理 4-1 设 X 表示实际损失额，分布函数为 $F(x)$。若保单规定了免赔额为 d，保单限额为 l 以及赔付比例为 α，则平均理赔额为：

$$E(Y)=\frac{\alpha[E(X\wedge l)-E(X\wedge d)]}{1-F(d)} \tag{4.2.2}$$

证明： 不妨设 X 为连续随机变量。对于每次损失额 X，设随机变量 Y^* 如下：

$$Y^*=\begin{cases}0, & X\leqslant d\\ \alpha(X-d), & d<X<l\\ \alpha(l-d), & X\geqslant l\end{cases}$$

则理赔额 $Y=Y^*\mid(X>d)$，由于

$$E(Y^*)=\int_d^l[\alpha(x-d)]f(x)dx+\alpha(l-d)[1-F(l)]$$

$$=\alpha\int_0^l xf(x)dx-\alpha\int_0^d xf(x)dx-\alpha d[F(l)-F(d)]$$

$$+\alpha(l-d)[1-F(l)]$$

$$=\alpha\left\{\left[\int_0^l xf(x)dx+l[1-F(l)]\right]-\left[\int_0^d xf(x)dx+d[1-F(d)]\right]\right\}$$

$$=\alpha[E(X\wedge l)-E(X\wedge d)]$$

所以，
$$E(Y)=\frac{E(Y^{*})}{1-F(d)}=\frac{\alpha[E(X\wedge l)-E(X\wedge d)]}{1-F(d)} \tag{4.2.3}$$

■

【例 4－3】 设某险种保单损失额 X 的概率密度函数为 $f(x)=0.04xe^{-0.2x}$，$x>0$，保单约定免赔额为 5 个单位，保单限额为 25 个单位，赔付比例为 80%。问：

（1）保单发生索赔的概率是多少？

（2）理赔额 Y 的期望是多少？

（3）当免赔额从 5 个单位提高到 10 个单位时，平均理赔额将会发生什么变化？

解：（1）由已知 $F(x)=\int_0^x f(t)dt=\int_0^x 0.04te^{-0.2t}dt=1-\frac{x+5}{5}e^{-\frac{x}{5}}$，可得：

$P(X>5)=1-F(5)=2e^{-1}=73.58\%$

即保单发生索赔的概率为 73.58%；

（2）由式（4.2.2）可知：

$$E(Y)=\frac{0.8\times[E(X\wedge 25)-E(X\wedge 5)]}{1-F(5)}$$

而
$$\begin{aligned}E(X\wedge 25)&=\int_0^{25}0.04x^2e^{-0.2x}dx+25[1-F(25)]\\&=0.04\int_0^{25}x^2e^{-0.2x}dx+25\times 6e^{-5}\\&=10-35e^{-5}=9.7642\end{aligned}$$

$$\begin{aligned}E(X\wedge 5)&=\int_0^{5}0.04x^2e^{-0.2x}dx+5[1-F(5)]\\&=0.04\int_0^{5}x^2e^{-0.2x}dx+5\times 2e^{-1}\\&=10-15e^{-1}=4.4818\end{aligned}$$

所以，
$$\begin{aligned}E(Y)&=\frac{0.8[E(X\wedge 25)-E(X\wedge 5)]}{1-F(5)}=\frac{0.8(9.7642-4.4818)}{0.7358}\\&=5.7433\end{aligned}$$

（3）因为
$$\begin{aligned}E(X\wedge 10)&=\int_0^{10}0.04x^2e^{-0.2x}dx+10[1-F(10)]\\&=0.04\int_0^{10}x^2e^{-0.2x}dx+10\times 3e^{-2}\\&=10-20e^{-2}=7.2933\end{aligned}$$

所以提高了免赔额后的平均理赔额为：

$$\begin{aligned}E(Y')&=\frac{0.8\times[E(X\wedge 25)-E(X\wedge 10)]}{1-F(10)}=\frac{0.8(9.7642-7.2933)}{3e^{-2}}\\&=4.8687\end{aligned}$$

平均免赔额下降的比例为 $\frac{E(Y)-E(Y')}{E(Y)}=15.23\%$。■

4.2.4 通货膨胀对理赔额分布的影响

在拟合损失分布时，我们所使用的经验数据来自过去某一时期内的损失额或理赔额，而我们所关心的是当前或未来某一时期内的损失额或理赔额的状况。随着时间的推移，同一险种的损失额自然会发生变化，其中通货膨胀（以下简称“通胀”）效应就是经常发生的一种。下面分两种情况讨论通胀效应问题。

1. 通胀率是确定的实数。设 X 表示某险种今年的每次实际损失额，分布函数为 $F_X(x)$，预计通胀率为 r，Z 为该险种明年的预计损失额，则 $Z=(1+r)X$，其分布函数为：$F_Z(z)=P[(1+r)X\leqslant z]=F\left(\frac{z}{1+r}\right)$

密度函数为：

$$f_Z(z)=\frac{1}{1+r}f\left(\frac{z}{1+r}\right)$$

且有

$$E(Z)=(1+r)E(X),\ Var(Z)=(1+r)^2Var(X)$$

下面我们来研究预计理赔额的分布。假设免赔额 d、保单限额 l 以及赔付比例 α 在通胀前后保持不变，设 X 为损失额，定义 Y^* 如下：

$$Y^*=\begin{cases}0, & X\leqslant \frac{d}{1+r}\\ \alpha[(1+r)X-d], & \frac{d}{1+r}<X<\frac{l}{1+r}\\ \alpha(l-d), & X\geqslant \frac{l}{1+r}\end{cases}$$

则通胀后预计的每次理赔额为 $Y=Y^*\mid\left(X>\frac{d}{1+r}\right)$，类似于定理 4－1 的证明，可以计算通胀后的平均理赔额为：

$$E(Y)=\frac{\alpha(1+r)\{E[X\wedge(l/(1+r))]-E[X\wedge(d/(1+r))]\}}{1-F_X\left(\frac{d}{1+r}\right)} \tag{4.2.4}$$

【例 4－4】 假设某险种 2009 年的每次实际损失额 X 服从离散分布，$P(X=1\,000k)=1/6,k=1,\cdots,6$。保单约定每次损失的免赔额为 2 000 元。假设从 2009 年到 2010 年的通胀率为 4% 且免赔额保持不变，求 2010 年的平均理赔额。与 2009 年相比，平均理赔额发生了什么变化？每次损失的平均赔付额发生了怎样的变化？这说明什么？

解： 由已知

$$E(X) = \frac{1}{6}(1+2+3+4+5+6) \times 1\ 000 = \frac{21\ 000}{6}$$

$$E(X \wedge 2\ 000) = \frac{1}{6} \times 1\ 000 + \frac{5 \times 2\ 000}{6} = \frac{11\ 000}{6}$$

$$E\left(X \wedge \frac{2\ 000}{1.04}\right) = \frac{1}{6} \times 1\ 000 + \frac{5 \times 2\ 000}{6 \times 1.04} = \frac{11\ 040}{6 \times 1.04}$$

那么，由式（4.2.1）可知 2009 年的平均理赔额为：

$$E(Y) = \frac{E(X) - E(X \wedge 2\ 000)}{1 - F(2\ 000)} = \frac{10\ 000/6}{1 - 1/3} = 2\ 500$$

再由（4.2.4）可知 2010 年的平均理赔额为：

$$E(Y') = \frac{1.04 \times \left[E(X) - E\left(X \wedge \frac{2\ 000}{1.04}\right)\right]}{1 - F\left(\frac{2\ 000}{1.04}\right)} = \frac{1.04 \times \left(\frac{21\ 000}{6} - \frac{11\ 040}{6 \times 1.04}\right)}{1 - 1/6}$$

$$= 2\ 160$$

$E(Y')/E(Y) = 2\ 160/2\ 500 = 0.8640$，这说明平均理赔额下降了 13.60%。

2009 年每次损失事件的平均赔付额为：

$$E(Y^{*}) = E(X) - E(X \wedge 2\ 000) = 10\ 000/6$$

2010 年每次损失事件的平均赔付额为：

$$E(Y^{*\prime})1.04 \times \left[E(X) - E\left(X \wedge \frac{2\ 000}{1.04}\right)\right] = 1\ 800$$

所以，相比 2010 年比 2009 年每次损失的提高了$\frac{1\ 800}{1\ 000/6} - 1 = 8\%$。

这个例子说明，通胀会增加理赔的成本，特别是当保单约定了免赔额时，通胀的影响可能会被放大，这其中的主要原因是通胀前低于免赔额的损失在通胀后可能会引起赔付，因此每次损失的平均赔付额增加了。但是由于小额赔付增加，导致每次理赔事件的理赔额减少了。 ■

2. 通胀率是随机变量。在很多情况下，通胀率是不确定的。设 X 表示某险种今年的每次实际损失额，通货膨胀率随机变量为 C，随机变量 C 和 X 独立，Y 为该险种明年的预计损失额，则 $Y = CX$。

设 X 的分布函数为 $F_X(x, \theta)$，θ 为参数，C 的分布函数为 $F_C(c)$，密度为 $f_C(c)$。假设对应任意实数 c，X 的分布满足 $F_{cX}(x, \theta) = F_X(x, c\theta)$ ①。于是，

$$F_Y(y) = \int_0^{\infty} P(CX \leqslant y \mid C = c) f_C(c)\, dc$$

$$= \int_0^{\infty} F_{cX}(y, \theta) f_C(c)\, dc = \int_0^{\infty} F_X(y, c\theta) f_C(c)\, dc$$

---

① 若分布族 $\{F_X(x, \theta), \theta \in \Theta\}$ 满足 $F_{cX}(x, \theta) = F_X(x, c\theta)$，其中 c 是实数，$c\theta \in \Theta$ 则称该分布族为尺度不变分布族，θ 称为尺度参数。本书所涉及的分布族都是尺度不变分布族。

$$f_Y(y) = \int_0^{\infty} f_X(y, c\theta) f_C(c)\,dc$$

容易计算出，第二年损失额 Y 的期望和方差为：

$$E(Y) = E(CX) = E(C)E(X) \tag{4.2.5}$$

$$Var(Y) = Var(X)E(C^2) + (E[X])^2 Var(C) \tag{4.2.6}$$

公式（4.2.6）的证明请读者自己练习。

【例 4－5】 经验数据显示明年的通胀率在 2%～6%之间，而且低通胀率的可能性更大。如果某险种的实际损失额 X 服从均值为 1 万元、标准差为 500 元的伽玛分布，试预测明年损失额的均值与标准差。

解：由已知，不妨假设 C 的概率密度函数为：

$$f_C(c) = \frac{1}{ac},\ 1.02 \leqslant c \leqslant 1.06$$

其中 $a = \int_{1.02}^{1.06} \frac{1}{c}dc = \ln\left(\frac{1.06}{1.02}\right) = 0.038466$ 。这个概率密度函数满足低通货膨胀率的可能性更大的假设条件。经计算，有

$$E(C) = \int_{1.02}^{1.06} c \cdot \frac{1}{a} \cdot \frac{1}{c}dc = \frac{1.06 - 1.02}{a} = 1.0399$$

$$E(C^2) = \int_{1.02}^{1.06} c^2 \cdot \frac{1}{a} \cdot \frac{1}{c}dc = \frac{1.06^2 - 1.02^2}{2a} - 1.0815$$

则由公式（4.2.5）、（4.2.6）计算得到：

$$E(Y) = E(C)E(X) = 1.0399 \times 10\,000 = 10\,399$$

$$\begin{aligned} Var(Y) &= Var(X)E(C^2) + (E[X])^2 Var(C) \\ &= 250\,000 \times 1.0815 + 10\,000^2 \times (2.1629 - 1.0399^2) \\ &= 527.10^2 \end{aligned}$$

因此明年损失额的均值与标准差分别是 10 399 元和 527.10 元。 ■

§4.3 理赔次数的分布

形成保单总赔付额不确定性的另一个原因是理赔次数的不确定性。本节将讨论单个保单（也可称为**个体保单**）以及保单组合的理赔次数分布。

由于保单理赔次数取值都是非负整数的特点，这一节我们将讨论计数随机变量的分布。计数随机变量是仅在非负整数上有概率的离散随机变量，它既可以用来描述损失次数，也可以用来描述理赔次数。

对于离散随机变量，除了可以利用矩母函数得到其各阶矩外，更为简便的方法是使用概率母函数。

定义 4－4 离散随机变量 N 的概率分布函数为 $p_k = P(N = k)$，$k = 0$，

1，2，…，其概率母函数为 $P_N(t) = E(t^N) = \sum_{k=0}^{+\infty} p_k t^k$。

与矩母函数类似，由概率母函数也可以得到随机变量的矩：$P_N'(1) = E(N), P_N''(1) = E[N(N-1)]$，并且还可以由概率母函数得到随机变量取不同值的概率，因为

$$P_N^{(m)}(t) = E\left(\frac{d^m}{dt^m}t^N\right) = E[N(N-1)\cdots(N-m+1)t^{N-m}]$$

$$= \sum_{k=m}^{\infty} k(k-1)\cdots(k-m+1)t^{k-m}p_k$$

所以，$P_N^{(m)}(0) = m!p_m$，$m = 1,2,\cdots$。即 $p_m = \frac{P_N^{(m)}(0)}{m!}$，$m = 1,2,\cdots$。

另外，由定义，矩母函数和概率母函数还存在如下关系：

$$M_N(t) = E(e^{tN}) = P_N(e^t) \tag{4.3.1}$$

4.3.1　单个保单的理赔次数分布

1. 泊松分布。随机变量 N 服从参数为 λ 的泊松分布，记做 $N \sim P(\lambda)$，其概率分布函数为：$p_k = e^{-\lambda}\frac{\lambda^k}{k!}$，$k = 0, 1, 2, \cdots$ 容易计算得到，泊松分布的概率母函数、期望和方差为：

$$P_N(t) = \sum_{k=0}^{\infty} p_k t^k = e^{\lambda(t-1)} \tag{4.3.2}$$

$$E(N) = Var(N) = \lambda \tag{4.3.3}$$

我们经常使用的泊松分布其实是泊松过程的单位化，其背景如下：保单的损失次数是指在单位时间内保单发生损失的次数。对一张特定保单来说，保单的损失事故可看做是稀有事件，因此可以假定它的损失次数过程 $N(t)$（随机过程 $N(t)$ 表示区间 $[0, t]$ 内发生的损失次数）具有如下特性：

（1）当 $t = 0$ 时，保单损失次数为0，即 $N(0) = 0$。

（2）在 $[t, t+\Delta t]$ 内保单发生损失这一事件与时刻 t 以前的损失事件相互独立，而且发生的损失次数只与时间长度 Δt 有关，与时间的起始位置无关。也就是说，$N(t)$ 是一个平稳独立增量过程。

（3）在充分小的时间间隔 Δt 中，至多有一次损失，且发生一次损失的概率与此时间区间的长度有关，而发生两次或两次以上的损失概率是 Δt 的无穷小量。即

$$P(N(\Delta t) = 1) = \lambda\Delta t + o(\Delta t)$$

$$P(N(\Delta t) \geqslant 2) = o(\Delta t)$$

满足上述三个条件的随机过程称为"泊松过程"。在 $[0, t]$ 内保单发生 k 次事故的概率为：

$$P(N(t)=k)=e^{-\lambda t}\frac{(\lambda t)^k}{k!},k=0,1,2,\cdots \tag{4.3.4}$$

令 $t=1$，则单位时间内损失次数 N 的概率分布函数为：

$$p_k=P(N=k)=e^{-\lambda}\frac{\lambda^k}{k!} \tag{4.3.5}$$

显然，N 服从泊松分布。

泊松分布具有两个非常有用的性质：可加性和可分解性。可加性是指有限个独立泊松随机变量的和依然服从泊松分布。这一性质可以用下面的定理表述：

定理 4－2 设 N_1，N_2，$\cdots$，N_n 是独立的泊松随机变量，参数分别为 λ_1，λ_2，$\cdots$，λ_n，则$N=N_1+N_2+\cdots+N_n$服从泊松分布，参数为 $\lambda=\lambda_1+\lambda_2+\cdots+\lambda_n$。

证明： 首先，独立随机变量和的概率母函数具有和矩母函数性质 3 类似的性质，再由式（4.3.2）可以得到 N 的概率母函数为：

$$P_N(z)=\prod_{i=1}^{n}P_{N_i}(z)=\prod_{i=1}^{n}\exp(\lambda_i(z-1))=\exp(\sum_{i=1}^{n}\lambda_i(z-1))$$

故 N 服从泊松分布，参数为 $\lambda=\lambda_1+\lambda_2+\cdots+\lambda_n$。 ■

泊松分布的可加性可用来判断保单组合总损失次数的分布。假设一个保单组合有 m 份保单，其中每份保单发生损失的次数服从参数为 λ 的泊松分布，则根据泊松分布的可加性可知，这个保单组合发生总损失的次数服从参数为 $m\lambda$ 的泊松分布。

泊松分布另一个有用的性质是可分解性，这是一个与可加性对立的性质。假设损失事故可以分为 m 个不同类型 C_1，C_2，$\cdots$，C_m 且相互独立。记 E_i 表示第 i 类事故发生，$p_i=P(E_i)$ 表示第 i 类事故发生的概率，N_i 表示第 i 类事故发生的次数，$i=1$，2，$\cdots$，m，$N=N_1+N_2+\cdots+N_m$ 表示所有事故发生的次数。若损失事故发生的次数 N 服从泊松分布，下面的定理将证明第 i 类事故发生的次数服从参数为 λp_i 的泊松分布：

定理 4－3 若 N 服从参数为 λ 的泊松分布，则 N_1，N_2，$\cdots$，N_m 相互独立，且服从泊松分布，参数分别是 λp_i，$i=1$，2，$\cdots$，m。

证明： 为证明本定理，先引入多项分布的概念，设一个随机试验 E 可能产生 m 个不同的结果，出现第 i 个结果的概率记为 π_i，$\sum_{i=1}^{m}\pi_i=1$，重复此试验 n 次，并记 n 次试验中结果 i 出现的次数为 N_i，$i=1,2,\cdots,m$，$\sum_{i=1}^{m}N_i=n$，则

$$P\{N_1=n_1,N_2=n_2,\cdots,N_m=n_m\}=\frac{n!}{n_1!n_2!\cdots n_m!}\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_m^{n_m}$$

称随机向量（N_1，N_2，$\cdots$，N_m）为具有参数（n，π_1，$\cdots$，π_m）的多项分布，多

项分布可以看做是对二项分布的推广。

给定 $N=n$，$N_i\mid(N=n)$ 服从二项分布 $B(n,\ p_i)$，$(N_1,\ N_2,\ \cdots,\ N_m)\mid(N=n)$ 服从多项分布 $B(n,\ p_1,\ \cdots,\ p_m)$。因此，

$$
\begin{aligned}
P(N_1=n_1,\cdots,N_m=n_m) &= P(N_1=n_1,\cdots,N_m=n_m\mid N=n)P(N=n)\\
&= \frac{n!}{n_1!n_2!\cdots n_m!}p_1^{n_1}\cdots p_m^{n_m}\frac{e^{-\lambda}\lambda^n}{n!}\\
&= \prod_{j=1}^{m} e^{-\lambda p_j}\frac{(\lambda p_j)^{n_j}}{n_j!}
\end{aligned}
$$

其中，$n=n_1+n_2+\cdots+n_m$，又

$$
\begin{aligned}
P(N_j=n_j) &= \sum_{n=n_j}^{\infty} P(N_j=n_j\mid N=n)P(N=n)\\
&= \sum_{n=n_j}^{\infty}\binom{n}{n_j}p_j^{n_j}(1-p_j)^{n-n_j}\frac{e^{-\lambda}\lambda^n}{n!}\\
&= e^{-\lambda}\frac{(\lambda p_j)^{n_j}}{n_j!}e^{\lambda(1-p_j)} = e^{-\lambda p_j}\frac{(\lambda p_j)^{n_j}}{n_j!}
\end{aligned}
$$

因此，$N=(N_1,\ N_2,\ \cdots,\ N_m)$ 的联合概率分布函数等于 $N_1,\ N_2,\ \cdots,\ N_m$ 概率分布函数的乘积，从而证明了 $N_1,\ N_2,\ \cdots,\ N_m$ 是相互独立的随机变量。 ■

【例 4－6】 考虑某项医疗保险计划，每人看病获赔的次数服从均值为 2.3 的泊松分布，保险公司为减少赔付次数，从保单中去掉一个保险项目。根据历史数据，该项目损失事故发生的概率为 0.1。请问去掉这个项目后，每人看病后获赔次数的期望值是多少？

解： 设 N 表示病人总共获得赔偿的次数，N_1 表示被撤销项目的获赔次数，N_2 表示未被撤销项目的获赔次数。根据泊松分布的可分解性，N_1、N_2 均服从泊松分布，每人看病后获赔次数的期望值是 $E(N_2)=2.3\times0.9=2.07$。 ■

【例 4－7】 设 N 表示保单理赔次数，X 表示理赔额。N 服从均值为 10 的泊松分布，X 服从区间 $[0,20]$ 上的均匀分布，求发生两次理赔额超过 5 的理赔事件的概率？

解： 令 M 表示理赔额超过 5 的事故次数，因为

$$P(X>5)=\int_5^{20}\frac{1}{20}dx=0.75$$

由定理 4－3，M 服从参数为 $10\times0.75=7.5$ 的泊松分布，所以，

$$P(M=2)=\frac{7.5^2}{2!}e^{-7.5}=0.0156$$ ■

泊松分布的均值和方差相等，但在实际运用中，并不是所有险种的保单损失次数或理赔次数的均值都等于方差，在这种情况下，还需要考虑其他的计数分布，如负二项分布、二项分布等。

2. 负二项分布。随机变量 N 服从参数为 r 和 p 的负二项分布，记做 $N \sim NB(r,p)$，其概率分布函数为：

$$p_k = P(N = k) = \binom{r+k-1}{k} p^r q^k, 0 < p < 1, p + q = 1, k = 0, 1, 2, \cdots$$

其中二项式系数为 $\binom{x}{k} = \frac{x(x-1)\cdots(x-k+1)}{k!}$，$k$ 为整数，x 为任意正实数。当 $r = 1$ 时，称 N 服从参数为 p 的几何分布。

不难证明负二项分布的概率母函数为：

$$P(t) = \left(\frac{p}{1-qt}\right)^r, \quad t < \frac{1}{q} \tag{4.3.6}$$

由此得到负二项分布的均值和方差：

$$E(N) = \frac{rq}{p}, \quad Var(N) = \frac{rq}{p^2} \tag{4.3.7}$$

因为 $0 < p < 1$，所以负二项分布的均值小于方差。注意到泊松分布的均值和方差相等，这说明如果观测数据的均值小于方差时，负二项分布比泊松分布更合适。

若令 $p = \frac{1}{1+\beta}$，即 $\beta = \frac{1-p}{p}$，也可以说 N 服从参数为 r 和 β 的负二项分布，我们称之为“奇异负二项分布”，记做 $N \sim NB(r, \beta)$，此时其概率母函数为：

$$P(t) = [1 - \beta(t-1)]^{-r}, \quad t < 1 + \frac{1}{\beta} \tag{4.3.8}$$

均值和方差分别是：

$$E(N) = r\beta, Var(N) = r\beta(1+\beta) \tag{4.3.9}$$

3. 二项分布。随机变量 N 服从参数为 n 和 p 的二项分布，记做 $N \sim B(n,p)$，其概率分布函数为 $p_k = P(N = k) = \binom{n}{k} p^k q^{n-k}, 0 < p < 1, p + q = 1, k = 0, 1, 2, \cdots, n$。二项分布的概率母函数、均值和方差如下：

$$P(t) = (pt + q)^n \tag{4.3.10}$$

$$E(N) = np, Var(N) = npq \tag{4.3.11}$$

可以看出，二项分布的均值大于方差，因此适用于拟合样本均值大于样本方差的数据。除此之外，二项分布的取值范围有限，这使得它比较适合描述理赔次数有限的索赔情况，例如，意外伤害保险的出险次数，交通事故的发生次数。另外，如果认为随机变量超过某个值的概率非常小的话，那么也可以近似地认为这个随机变量服从二项分布。

4. 泊松分布与负二项分布的关系。在保险实务中，不同的保单类型或同一保单类型在不同保险期间的平均理赔次数是一个随机变量，也就是说，

如果理赔次数 N 服从均值为 λ 的泊松分布，参数 λ 是随机的，记做 Λ，设 Λ 的概率密度函数为 $u(\lambda)$、$\lambda>0$，也就是说，当 Λ 取定为某个值 λ 后，N 服从参数为 λ 的泊松分布，运用全概率法则，可得 N 的最终（无条件）分布：

$$P(N=n)=\int_0^\infty P(N=n\mid\Lambda=\lambda)u(\lambda)d\lambda$$
$$=\int_0^\infty \frac{\lambda^n e^{-\lambda}}{n!}u(\lambda)d\lambda$$

由（4.3.3）还可以得到 N 的无条件期望和方差，但是需要用到下面的命题：

命题 4－1　设 X 和 Y 是任意的随机变量，而且数学期望和方差都存在，则有：

$$E(X)=E[E(X\mid Y)] \tag{4.3.12}$$
$$Var(X)=E[Var(X\mid Y)]+Var[E(X\mid Y)] \tag{4.3.13}$$

大多数的概率统计教材中都有关于这个命题的证明，这里不再重复，请读者自行参阅。

有了命题4－1，再由式（4.3.3）可得：

$$E(N)=E[E(N\mid\Lambda)]=E[\Lambda] \tag{4.3.14}$$
$$Var(N)=E[Var(N\mid\Lambda)]+Var[E(N\mid\Lambda)]$$
$$=E(\Lambda)+Var(\Lambda) \tag{4.3.15}$$
$$M_N(t)=E(e^{tN})$$
$$=E[E(e^{tN}\mid\Lambda)]=E[e^{\Lambda(e^t-1)}]$$
$$=M_\Lambda(e^t-1) \tag{4.3.16}$$

因此，可以得到如下负二项分布与泊松分布关系的定理：

定理 4－4　若已知当 Λ 取定为某个值 λ 后，N 服从参数为 Λ 的泊松分布，而且 Λ 服从伽玛分布，参数为(α,θ)，则 N 的无条件分布为奇异负二项分布，参数分别为：

$$r=\alpha,\beta=\theta \tag{4.3.17}$$

证明：由 Λ 服从伽玛分布，参数为(α,θ) 可知其矩母函数为：

$$M_\Lambda(t)=(1-\theta t)^{-\alpha},0<t<\frac{1}{\theta}$$

将其代入式（4.3.16），有

$$M_N(t)=M_\Lambda(e^t-1)=[1-\theta(e^t-1)]^{-\alpha},0<t<\ln\left(1+\frac{1}{\theta}\right)$$

对照式（4.3.1）和（4.3.8）可知，N 服从参数为 (r,θ) 的负二项分布。

■

【例 4－8】　某保单承保了 n 个学生为期一周的团体意外伤害保险，保

险责任是死亡。假设这 n 个学生是独立同分布的风险个体，死亡率均为 p，求这个团体总理赔次数的分布？

解： 根据题意，设每个学生的理赔次数为 $N_i, i = 1,2,\cdots,n$，它服从参数为 p 的两点分布（贝努里分布），其概率母函数为 $P_N(t) = pt + q$；再设团体总理赔次数为 N，因为独立随机变量和的概率母函数是概率母函数的乘积，所以有 $P_N(t) = (P_{N_i}(t))^n = (pt + q)^n$。

由概率母函数的性质，可证明 $P(N = k) = \binom{n}{k}p^k q^{n-k}, 0 < p < 1, p + q = 1$，$k = 0,1,2,\cdots,n$，这说明团体总理赔次数服从参数为 n 和 p 的二项分布。

■

【例 4－9】 Daykin（1994）记录了英国某种综合汽车保险 1968 年的索赔情况。共计观察了 421240 张保单，具体的理赔次数记录列在表 4－1 中。试分析该组数据适用的分布和参数估计。

解： 经过简单的计算，可得这组观测的平均理赔次数是 0.13174，样本方差是 0.13825。因为两者比较接近，所以考虑泊松分布。泊松参数的估计值 $\hat{\lambda}$ 为平均理赔次数 0.13174。

为了对比不同分布的拟合效果，还可以考虑用负二项分布来拟合。对于负二项分布，若以 $\overline{N}$ 表示样本均值、s_N^2 表示样本方差，则由式（4.3.7）式可得参数估计：

$$\hat{p} = \frac{\overline{N}}{s_N^2} = 0.951$$

$$\hat{r} = \frac{\hat{p}}{1 - \hat{p}}\overline{N} = 2.555$$

表 4－1 对比了理赔次数的实际观察值、泊松分布的估计值和负二项分布的估计值。读者还可以用 χ^2 分布来检验拟合的情况。

表 4－1　例 4－9 的理赔次数记录及参数估计

理赔次数 k	观察记录	泊松分布 ($\hat{\lambda}$ =0.13174)	负二项分布 ($\hat{p}$ =0.951，$\hat{r}$ =2.555)
0	370 412	369 246	370 460
1	46 545	48 644	46 411
2	3 935	3 204	4 045
3	317	141	301
4	28	5	21
5	3	—	1

■

4.3.2　（a，b，0）分布类和（a，b，1）分布类

1.（a，b，0）分布类。

定义 4－5　计数随机变量 N 的概率分布函数为 p_k，若存在常数使得下式成立：

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k = 1,2,3,\cdots \tag{4.3.18}$$

则称随机变量 N 属于（a，b，0）分布类。（a，b，0）分布类是一个两参数分布族，参数分别是 a 和 b。

将泊松分布、负二项分布和二项分布的概率分布函数代入式（4.3.18）的左边，可以看出三个分布均满足该递推公式，如表 4－2 所示，由于几何分布是负二项分布参数 $r=1$ 时的特例，所以表 4－2 也给出了几何分布的参数值，同时表 4－2 也列出了这四个分布的 p_0 值。

表 4－2　（a，b，0）类的分布

分布	a	b	p_0
泊松分布	0	λ	$e^{-\lambda}$
负二项分布	$1-p$	$(r-1)(1-p)$	p^r
几何分布	$1-p$	0	p
二项分布	$-\frac{p}{q}$	$(n+1)\frac{p}{q}$	q^m

可以证明，也只有这些分布满足上述的递推公式。递推公式（4.3.18）也可以表示为：

$$k\frac{p_k}{p_{k-1}}=ak+b\,,\qquad k=1,2,3,\cdots \tag{4.3.19}$$

由式（4.3.19）可以看出，函数 $k\frac{p_k}{p_{k-1}}$ 是 k 的线性函数，它的图形是一条斜率为 a、截距为 b 的直线。

由表 4－2 可以看出，泊松分布、负二项分布和二项分布的斜率 a 分别是 0、正数和负数，这一特点可以帮助我们选择合适的理赔次数分布。首先，我们可以按照下面的近似公式画出关于 k 的图形：

$$k\frac{\hat{p}_k}{\hat{p}_k-1}=k\frac{n_k}{n_k-1}\ \text{（}n_k\text{ 表示发生 }k\text{ 次事故的保单数）} \tag{4.3.20}$$

若由观测值画出的图形近似是一条直线，那么大致可判断其属于（a，b，0）分布族，直线的斜率表示适用的模型。需要注意的是，如果样本数据中出现了某个 n_k 为 0，那么这种方法就不太适用。

2.（a，b，1）分布类。在保险实践中，我们发现有时候（a，b，0）分布不能充分地反映经验数据的特征，特别是端点的特征，而非寿险业务中免赔额和保单限额的存在又使得理赔次数的分布很容易出现端点特别是零点的概率异常。理赔次数在零点的概率表示保单在观察期内没有发生索赔的概率，由于保险事故发生的比率一般都很低，因此理赔次数在零点有较大的概率值。考虑到要更为准确地拟合零点的概率值，我们有必要对（a，b，0）分布族在零点的值作调整。

定义 4－6　设计数随机变量 N 的概率分布函数为 p_k，$k\geqslant 0$。若存在常数使得下式成立：

$$\frac{p_k}{p_{k-1}}=a+\frac{b}{k}\,,\qquad k=2,3,4,\cdots \tag{4.3.21}$$

则称随机变量 N 属于$(a,b,1)$分布类。

比较$(a, b, 0)$分布类和$(a, b, 1)$分布类的定义，我们发现，$(a, b, 1)$分布类与$(a,b,0)$分布类唯一的区别是递推从 p_1 而非从 p_0 开始，即$(a, b, 1)$分布类只在从 $k = 1$ 到 $k = \infty$ 的概率值之间存在递推关系，而$(a, b, 0)$分布类却在从 $k = 0$ 到 $k = \infty$ 的概率值之间存在递推关系。当然这种递推关系可能不同，但却只相差一个常数，也就是说两类分布在从 $k = 1$ 到 $k = \infty$ 的概率值之间存在倍数关系，这一结论可以在定理 4－5 的证明过程中得到。

从另一角度看，$(a,b,1)$分布类与$(a,b,0)$分布类的本质区别就是分布在零点的取值 p_0，这个值使得$(a,b,1)$分布类成为三参数的分布：a,b,p_0，而$(a,b,0)$分布类只是含两个参数 a 和 b 的分布类。

下面我们就 $p_0 = 0$ 和 $p_0 > 0$ 两种情况来分别讨论：

- 当 $p_0 = 0$ 时，随机变量 N 的取值从 $k = 1$ 开始，从概率分布函数的图形上看，相当于在$(a,b,0)$类分布的基础上再截去零点的值，这一类分布称为“零点截断分布”（zero-truncation）或“ZT 分布”，其概率分布函数用 p_k^T 表示。零点截断分布包括零点截断泊松分布、零点截断负二项分布以及零点截断二项分布。

- 当 $p_0 > 0$ 时，随机变量 N 在各点的取值是由$(a,b,0)$类修正得到的，因此称之为“零点修正分布”（zero-modification）或“ZM 分布”，其概率分布函数用 p_k^M 表示。可以看出，$(a, b, 0)$分布类属于$(a, b, 1)$分布类中的零点修正分布类，是零点截断分布类和恒为 0 的随机变量的混合，这一结论由定理 4－4 给出。

定理 4－4　任一零点修正分布是零点截断分布和恒为 0 的随机变量的混合。

证明：对于任一零点修正分布，其概率分布函数为 p_k^M，$k = 0,1,2,\cdots$，令

$$p_0^T = 0 \,,\, p_k^T = \frac{p_k^M}{1 - p_0^M}, \qquad k = 1, 2, \cdots$$

那么有

$$p_k^M = (1 - p_0^M)p_k^T + p_0^M d_k, \qquad k = 0, 1, 2, \cdots$$

其中，$d_0 = 1$，$d_k = 0$，$k = 1,2,\cdots$。d_k 正是恒为 0 的随机变量的概率分布函数，而 p_k^T 对应的随机变量是零点截断分布。 ■

定理 4－5　任一零点修正分布是（a，b，0）分布和恒为 0 的随机变量的混合。

证明：设$(a, b, 0)$分布的概率分布函数为 p_k，概率母函数为 $P(t)$；对应的零点修正分布的概率分布函数为 p_k^M，概率母函数为 $P^M(t)$。比较

(4.3.18) 和 (4.3.21) 知存在常数 c，使得

$$p_k^M = cp_k, \quad k = 1, 2, 3, \cdots, \tag{4.3.22}$$

则有

$$\begin{aligned} P^M(t) &= p_0^M + \sum_{k=1}^{\infty} p_k^M t^k = p_0^M + c\sum_{k=1}^{\infty} p_k t^k \\ &= p_0^M + c[P(t) - p_0] \end{aligned} \tag{4.3.23}$$

再把 $P^M(1) = P(1) = 1$ 代入式 (4.3.22)，有 $1 = p_0^M + c(1 - p_0)$，所以可得 $c = \dfrac{1 - p_0^M}{1 - p_0}$。因此有

$$\begin{aligned} P^M(t) &= p_0^M + \frac{1 - p_0^M}{1 - p_0}[P(t) - p_0] \\ &= \left(1 - \frac{1 - p_0^M}{1 - p_0}\right) \times 1 + \frac{1 - p_0^M}{1 - p_0}P(t) \end{aligned} \tag{4.3.24}$$

式 (4.3.24) 中，1 正是定理 4－4 中概率分布函数为 d_k 的随机变量的概率母函数，再由概率母函数的唯一性，我们就证明了任一零点修正分布是 $(a,b,0)$ 分布和恒为 0 的随机变量的分布的混合。 ■

由定理 4－5 可知，对于零点修正的 $(a,b,1)$ 类分布，其概率分布函数与对应的 $(a,b,1)$ 类分布的概率分布函数之间存在如下倍数关系：

$$p_k^M = \frac{1 - p_0^M}{1 - p_0}p_k, \quad k = 1,2,3,\cdots \tag{4.3.25}$$

【例 4－10】 假设某车队理赔次数服从零修正泊松分布，泊松参数为 4。经验数据表明，发生索赔的概率为 40%，在这种零点修正的条件下，试给出车队发生 3 次索赔的概率。

解： 首先由题意可知 $p_0^M = 1 - 0.4 = 0.6$，又 $p_0 = e^{-4}$，$p_3 = \dfrac{32}{3}e^{-4} = 0.195367$，那么由式 (4.3.24) 可知：

$$p_3^M = \frac{1 - 0.6}{1 - e^{-4}}p_3 = \frac{0.4}{1 - e^{-4}} \times \frac{32}{3}e^{-4} = 0.079605$$

这说明，车队发生 3 次索赔的概率已经由 19.54% 修正为 7.96%。 ■

事实上，我们还可以定义更一般的 (a,b,k) 分布类。

3. (a, b, k) 分布类。称计数随机变量 N 属于 (a, b, k) 分布类，若存在常数使得下式成立：

$$\frac{p_j}{p_{j-1}} = a + \frac{b}{j}, \quad j = k + 1, k + 2, \cdots$$

$$p_j = P(N = j), \quad j = 0, 1, 2, \cdots$$

(a, b, k) 分布是一类具有 $2 + k$ 个参数的分布。

在拟合理赔次数的经验数据时，使用 $(a, b, 0)$ 分布、$(a, b, 1)$ 分布还是 (a, b, k) 分布，要依实际的数据而定。当假设经验数据属于 $(a ,b, 0)$ 分

布时，我们只需估计两个参数：a 和 b；但当这种假设没有通过假设检验时，我们可以放宽假设为$(a, b, 1)$分布，这样就需要估计三个参数：a、b 和 p_0；进一步，如果$(a, b, 1)$分布的假设也没有通过假设检验，我们只能假设数据属于$(a, b, 2)$分布甚至$(a, b, 3)$分布，等等。如此，我们就需要估计更多的参数。

4.3.3 保单组合的理赔次数分布

在保险实践中，保险公司获得的是同险种多张保单的总损失或理赔次数，而不只是单个保单的损失次数。因此，有必要研究从保单组合中随意抽取一份保单，该保单的损失次数的问题，从而能够研究保单组合损失次数分布的问题。

对于保单组合，可以分为同质性和非同质性两种情况来考虑。同质性是指所有的保单相互独立，且都有相同的风险水平，即各保单的损失额分布相同，损失次数的分布也相同。在这种情况下，任意一份保单的风险水平都代表了整体的风险水平。例如，若假设 n 份个体保单的损失次数都服从参数为 λ 的泊松分布，则保单组合的损失次数分布也服从泊松分布，其参数为 $n\lambda$ 。但是，事实上，没有任何两份保单具有相同的风险水平。为了保证保单组合的同质性，保险公司通常根据某些风险分级变量取不同的值将不同的被保险人进行分组，如性别、年龄、地域等。风险分级变量的选取虽然尽量使得同一组内的被保险人具有相似的风险水平，但是由于经济的、社会的原因，分级变量不可避免地存在一些缺陷。因此，不可能存在完全同质的保单组合。对于完全非同质性的保单组合，可以考虑混合损失次数模型。

为了简化问题，我们假设所有保单的损失额分布都相同，对于个体保单，我们假定损失次数服从泊松分布，分布的参数 λ 反映了其风险水平的高低。在同一个保单组合中，每份保单的参数 λ 是有差异的。因此，为了要描述一个非同质性的保单组合的损失次数，就首先需要确定这个保单组合中泊松参数 λ 的变化规律。对于规模较小的保单组合，可以考虑分组的方法，按风险水平的高低将被保险人分成几组，每个泊松变量取有限的值。例如，用大、中、小三种取值分别代表高、中、低三种风险水平。对于规模较大的保单组合，可以假设其中的泊松参数服从连续分布。以 $u(\lambda)$ 表示 Λ 的密度函数，通常称为“结构函数”。那么，从保单组合中随机抽取一份保单的损失次数分布为：

$$P(N = k) = \int_0^\infty \frac{e^{-\lambda}\lambda^k}{k!}u(\lambda)d\lambda , \quad k = 0,1,2,\cdots \tag{4.3.26}$$

形如式（4.3.26）的分布称为**混合泊松分布**。由条件期望公式，可以计算得到：

$$E(N) = E[E(N|\Lambda)] = E(\Lambda)$$

$$Var(N) = Var(\Lambda) + E(\Lambda) \tag{4.3.27}$$

由此我们可以看出，混合泊松分布的一个显著特点是方差大于均值。我们知道，泊松分布的均值等于方差，因此当保单组合损失次数观察值的样本方差大于其均值时，可以判断保单组合中存在某种程度的非同质性，而且方差越大于均值，这种非同质性就越严重。这从公式（4.3.27）很容易看出，$Var(\Lambda)$ 正是反映这种保单组合中非同质性程度的结构函数的方差。下面我们分别讨论三种常见的结构函数：

1. 离散结构函数。假设保单组合由 n 种不同的风险水平构成，泊松参数 Λ 取值于 $\{\lambda_1,\cdots,\lambda_n\}$ ，$\lambda_1 < \cdots < \lambda_n$ ，我们把这种模型称为"n 元结构模型"。设 $P(\Lambda = \lambda_i) = a_i$ ，$i = 0,1,2,\cdots,n$ ，当 $\Lambda = \lambda_i$ 时，保单的损失次数服从参数为 λ_i 的泊松分布，则从保单组合中任意抽取一份保单，其理赔次数的概率分布函数为：

$$P(N = k) = \sum_{i=1}^{n} P(N = k|\Lambda = \lambda_i)P(\Lambda = \lambda_i)$$

$$= \sum_{i=1}^{n} a_i \frac{\lambda_i^k}{k!}e^{-\lambda_i},\ k = 0, 1, 2, \cdots$$

通常使用的是二元结构模型，即 $n = 2$ 。设保单组合由两类风险水平组成，其中低风险（泊松分布参数 λ_1 ）的比例为 a_1 ，高风险（泊松分布参数 λ_2 ）的比例为 $a_2 = 1 - a_1$ ，于是，

$$P(N = k) = a_1\frac{\lambda_1^k}{k!}e^{-\lambda_1} + a_2\frac{\lambda_2^k}{k!}e^{-\lambda_2},\ \lambda_1 > 0,\ \lambda_2 > 0,\ k = 0, 1, 2, \cdots \tag{4.3.28}$$

经计算得到它的均值、方差和三阶中心矩为：

$$E(N) = a_1\lambda_1 + a_2\lambda_2$$

$$Var(N) = a_1\lambda_1^2 + a_1\lambda_1 + a_2\lambda_2^2 + a_2\lambda_2 - (a_1\lambda_1 + a_2\lambda_2)^2$$

$$\mu_3 = \alpha_3 - 3m\alpha_2 + 2m^3,\ \alpha_3 = a_1\lambda_1^3 + a_2\lambda_2^3 + 3(a_1\lambda_1^2 + a_2\lambda_2^2) + a_1\lambda_1 + a_2\lambda_2$$

【例 4-11】 某保单组合中，被保险的汽车司机每年发生事故次数服从泊松分布，假设把司机根据发生事故次数的多少分为两类，其中驾驶记录良好的一类泊松参数为 0.11，驾驶记录较差的另一类泊松参数为 2.3，如果保单组合中有 10% 的司机行车记录较差，求任意一个司机每年发生 2 次事故的概率。

解： 由已知 $\lambda_1 = 0.11$ ，$\lambda_2 = 2.3$ ，$a_1 = 0.9$ ，$a_2 = 0.1$ ，那么由式（4.3.28）可知司机发生事故次数的概率分布为：

$$P(N = k) = \alpha_1\frac{\lambda_1^k}{k!}e^{-\lambda_1} + \alpha_2\frac{\lambda_2^k}{k!}e^{-\lambda_2} = 0.9 \times \frac{0.11^k}{k!}e^{-0.11} + 0.1 \times \frac{2.3^k}{k!}e^{-2.3}$$

因此，

$$P(N=2)=0.9\times\frac{0.11^2}{2!}e^{-0.11}+0.1\times\frac{2.3^2}{2!}e^{-2.3}=3.14\%$$

■

2. 伽玛结构函数。假设个体保单的损失次数服从泊松分布，其中参数 Λ 随保单不同而不同；再假设 Λ 服从伽玛分布，参数为$(\alpha,\ \theta)$。根据定理 4－4 知，N 服从负二项分布。

3. 逆高斯结构函数。当实际数据的尾部较厚时，伽玛分布作为结构函数不是特别理想。逆高斯分布是一个比较理想的选择，它的概率密度函数为：

$$u(\lambda)=r(2\pi\beta\lambda^3)^{-\frac{1}{2}}\exp\left\{-\frac{(\lambda-r)^2}{2\beta\lambda}\right\},r>0,\qquad\beta>0,\lambda>0$$

保单组合中任意一份保单损失次数分布为泊松—逆高斯分布：

$$p_0=\exp\left\{-\frac{r}{\beta}[(1+2\beta)^{\frac{1}{2}}-1]\right\}$$

$$p_k=p_0\cdot\frac{r^k}{k!}\left[\sum_{m=0}^{k-1}\frac{(k+m-1)!}{(k-1)!m!}\cdot\left(\frac{\beta}{2r}\right)^m\cdot(1+2\beta)^{-\frac{k+m}{2}}\right],k=1,2,3,\cdots$$

其均值、方差和偏度系数 γ 为：

$$E(N)=r$$

$$Var(N)=r(1+\beta)$$

$$\gamma=\sigma^{-3}\left\{3\sigma^2-2\mu+3\frac{(\sigma^2-\mu)^2}{\mu}\right\}$$

与负二项式分布相比，当两者都具有相同的均值和方差时，泊松－逆高斯分布的偏度系数较大，因此它的尾部较厚。

4.3.4 相关性保单组合的理赔次数分布

有些情况下，一次保险事故的发生可能导致多份保单同时发生损失。这类险种的保单，我们通常称为**相关性保单**。例如，在火灾险中，一次大火可能导致多户在同一家保险公司投保火灾保险的家庭要求索赔；一次飞机事故，会导致所有购买保险的保单产生索赔。对于相关性保单组合，我们通常使用复合分布来描述。

设 N 表示保单组合在单位时间内发生事故的次数，其概率分布函数为 $p_n=P(N=n)$，概率母函数为 $P_N(t)$。M_i 表示第 i 次事故中产生的索赔数，M_i，$i=1,2,\cdots,N$ 为独立同分布的随机变量，且与事故次数 N 独立，M_i 的概率分布函数为 $p_m=P(M_i=m)$，概率母函数为 $P_M(t)$，那么在单位时间内保单组合发生的总理赔次数为：

$$S=M_1+\cdots+M_N$$

因为 S 由随机变量 M_i 与 N 复合而成，因此称其为**复合随机变量**。利用独立随机变量和的性质，可以得到 S 的概率分布函数为：

$$
\begin{aligned}
P(S=k) &= \sum_{n=0}^{\infty} P(N=n)P(S=k \mid N=n) \\
&= \sum_{n=0}^{\infty} P(N=n)P(M_1+M_2+\cdots+M_n=k) \\
&= \sum_{n=0}^{\infty} p_n \cdot p_m^{*n}(k)
\end{aligned}
$$

其中，$p_m^{*n}(k)=P(M_1+M_2+\cdots+M_N=k)$ 称为 p_m 的 n 重卷积。S 的概率母函数为：

$$
\begin{aligned}
P_S(t) &= \sum_{k=0}^{\infty} P(S=k)t^k = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} P(N=n)P(S=k \mid N=n)t^k \\
&= \sum_{n=0}^{\infty} P(N=n)\sum_{k=0}^{\infty} P(M_1+M_2+\cdots+M_N=k \mid N=n)t^k \\
&= \sum_{n=0}^{\infty} P(N=n)\sum_{k=0}^{\infty} P(M_1+M_2+\cdots+M_n=k)t^k \\
&= \sum_{n=0}^{\infty} P(N=n)P_{M_1+M_2+\cdots+M_n}(t) \\
&= \sum_{n=0}^{\infty} P(N=n)[P_M(t)]^n \\
&= P_N(P_M(t))
\end{aligned}
$$

由命题 4.1 可以得到 S 的均值和方差为：

$$E(S)=E(N)E(M) \tag{4.3.29}$$

$$Var(S)=E(N)Var(M)+E(M)^2Var(N) \tag{4.3.30}$$

【例 4－12】 某航空公司每月从城市 A 到城市 B 的航线有 70 个航班，假设每个航班被取消的可能性为 2%，并且每次飞行发生事故的概率为 0.00001。已知运营的飞机有 200 个座位，每次飞行乘客的就座率为 90%，机组人员为 6 名。假设机上所有人员都购买了保险并且飞机发生事故将致所有人员死亡从而引起索赔。求此航线每月理赔次数的期望和方差。

解： N 表示每个月出行的航班数，$N \sim B(n_1, p)$，$n_1=70$，$p=0.98$。

$$E(N)=n_1p=68.6,\ Var(N)=70\times0.98\times0.02=1.372$$

P 表示飞机上的人员数，M 表示飞机乘客数，$M \sim B(n_2,p)$，$n_2=200$，$p=0.9$，则 $P=6+M$

$$E(P)=6+200\times0.9=186$$

$$Var(P)=200\times0.9\times0.1=18$$

令 O 表示发生事故的死亡人数，则

$$E(O)=0.00186$$

$$E(O^2)=0.00001E(P^2)=0.34614$$

$$Var(O)=0.34167-0.00186^2=0.346137$$

令 S 表示下个月此航线的理赔次数，则

$$S=O_1+O_2+\cdots+O_N$$

$$E(S) = E(N)E(O) = 68.6 \times 0.00186 = 0.1276$$

$$Var(S) = E(N)Var(O) + E(O)^2 Var(N)$$
$$= 68.6 \times 0.346137 + 0.00186^2 \times 1.372 = 23.7450$$

■

4.3.5 免赔额对理赔次数的影响

当保单规定免赔额时，理赔额不等于损失额，索赔的次数也不等于损失次数。但是，两者又是密切相关的。同上节一样，我们将在损失次数的基础上来讨论理赔次数的分布。设 X 表示实际发生的损失，$f_X(x)$ 为其密度函数，d 表示免赔额，v 表示保单发生索赔的概率，则 $v = P(X > d)$，Y 表示理赔额，$Y = (X - d) \mid (X > d)$。

令 N 表示保单组合单位时间内发生损失的次数，$I_j = 0$ 表示第 j 个损失事故不会引起赔偿，$I_j = 1$ 表示第 j 个损失事故会引起赔偿，$j = 1, 2, \cdots, N$，则 $P(I_j = 1) = v$。再令 N^* 表示免赔额为 d 时的理赔次数，则

$$N^* = I_1 + I_2 + \cdots + I_N \tag{4.3.31}$$

若 $I_1, \cdots, I_N$ 相互独立，且与 N 独立，则 N^* 是一个复合分布，其概率母函数为：

$$P_{N^*}(t) = P_N[P_I(t)] = P_N[1 + v(t - 1)] \tag{4.3.32}$$

其中 $P_I(t) = E(t^I) = t^0 P(I = 0) + t^1 P(I = 1) = 1 - v + tv = 1 + v(t - 1)$。

【例 4－13】 假设某险种的损失额服从帕累托分布，$\alpha = 3$，$\theta = 1\ 000$，免赔额为 250 元。假设损失次数服从奇异负二项分布，$r = 2$，$\beta = 3$，求理赔次数的分布。

解： 由已知损失额 X 的分布函数为 $F(x) = 1 - \left(\dfrac{1\ 000}{x + 1\ 000}\right)^3$，因此索赔的概率为 $v = P(X > 250) = 0.512$。由式（4.3.32），有

$$P_{N^*}(t) = \{1 - 3[[1 + 0.512(t - 1)] - 1]\}^{-2} = [1 - 1.536(t - 1)]^{-2}$$

所以理赔次数服从奇异负二项分布，参数分别为 2 和 1.536，这说明保单增加免赔额条款后，理赔次数的均值由 6 下降到 3.072，标准差由 4.899 下降到 2.7910。 ■

类似地，我们也可以分析免赔额提高后理赔次数分布的变化情况。N' 表示把免赔额提高到 $d' > d$ 后的理赔次数，设 v 表示在免赔额提高后，以前的索赔事件能够继续获得赔偿的比例，则 $v' = \dfrac{1 - F_X(d')}{1 - F_X(d)}$。令 $I'_j = 1$ 表示继续获得赔偿，$I'_j = 0$ 表示不能继续获得赔偿，$j = 1, 2, \cdots, N^*$，$P(I'_j = 1) = v'$，则

$$N' = I_1 + I_2 + \cdots + I_{N^*}$$

若假设 $I_1, I_2, \cdots, I_{N^*}$ 相互独立，且与 N^* 独立，则 N' 仍是复合分布，其概率

母函数为：

$$P_{N'}(t) = P_{N^*}[P_{I'}(t)]$$

把 $P_{I'}(t) = 1 + v'(t-1)$ 代入式（4.3.32），则有

$$P_{N'}(t) = P_N[1 + v[1 + v'(t-1) - 1]] = P_N[1 + vv'(t-1)]$$

注意，若免赔额降低，则 $v' = \dfrac{1 - F_X(d')}{1 - F_X(d)} > 1$，此时 N^* 的参数可能超出频率分布的范围，我们不考虑这种情形。

【例 4－14】 承例 4－13，在其他条件不变的情况下，若把免赔额提高到 300 元，考虑理赔次数的分布。

解： 首先由 $v' = \dfrac{1 - F(300)}{1 - F(250)} = \dfrac{\left(\dfrac{1\,000}{1\,300}\right)^3}{0.512} = 0.8890$，可知

$$1 + vv'(t-1) = 1 + 0.512 \times 0.8990(t-1) = 0.4552t + 0.5448$$

所以，

$$P_{N'}(t) = [1 - 3(0.4552t + 0.5448 - 1)]^{-2} = [1 - 1.3656(t-1)]^{-2}$$

即理赔次数仍服从奇异负二项分布，参数分别为 2 和 1.3656。■

习　题

1. 假设某车险实际损失额 X 的分布函数为：

$$F(x) = 1 - 0.9e^{-0.02x} - 0.1e^{-0.001x}, x \geqslant 0$$

假设保单规定了保单限额为 5 000 元，求平均理赔额。

2. 某班级的数学期末考试成绩服从期望为 θ、标准差为 8 的正态分布。θ 是一个服从期望为 75、标准差为 6 的正态分布的随机变量。数学老师制定了如下的奖励规则：老师每年随机选择一个学生，如果其数学考试成绩高于 65 分，那么这个学生就可获得奖学金，奖学金数额正好就是考试分数。考虑在获得奖学金的条件下，奖学金金额小于 90 的概率。

3. 一个离散概率分布有如下性质：（1）$p_k = c\,(1 + 1/k)p_{k-1}$，$k = 1, 2, \cdots$；（2）$p_0 = 0.5$。计算 c。

4. 某路公交车到站的数量服从每小时 20 辆的泊松分布，其中 25% 的车是快车，75% 的车是慢车；另外，公交车到站的类型与数量互相独立。某人乘坐公交车上班，从车站到工作单位，快车需要 16 分钟，慢车需要 28 分钟。通常他总是乘坐最先到站的任何一种公交车，而他的同事则总是乘坐最先到站的快车。假设此人和他的同事在同一个车站候车，计算在慢车先到达的情况下，此人比其同事先到达单位的概率。

5. 损失随机变量 X 服从混合分布：（1）80% 的点服从 $\alpha = 2$，$\theta = 100$ 的帕累托分布；（2）20% 的点服从 $\alpha = 4$，$\theta = 3\,000$ 的帕累托分布。计算

$P(X \leqslant 200)$。

6. 已知某社区的每个人在一年中患普通感冒的次数服从泊松分布，患感冒的次数与个体的年龄和吸烟状况有关，具体数据见表 4－3。已知某人在一年里得了 3 次感冒，计算他是一个成年吸烟者的条件概率。

表 4－3

	占总人口的比例	患感冒次数的期望值
儿童	0.3	3
成年非吸烟者	0.6	1
成年吸烟者	0.1	4

7. 损失随机变量 X 的分布函数如下：

$$F(x) = \begin{cases} \left(\dfrac{x}{100}\right)^2, & 0 \leqslant x \leqslant 100 \\ 1, & x > 100 \end{cases}$$

一份保单约定每次赔付的免赔额是 20，保单限额是 60，赔付比例是 80%。计算每次赔付额的平均值。

8. 一个游戏的支付次数和支付额度的规律如下：支付次数服从每小时 5 次的泊松分布，每次的支付额可为 1,2,3,… 并且不设上限，每次支付等于 i 的可能性是 $1/2^i$，并且每次支付均独立。计算在给定的 20 分钟内没有任何支付的概率。

9. 某人早上 6:15 到达火车站，根据他的经验，火车在 7 点之前会以每半小时 1 列车的速度到站；而 7 点之后，则会以每半小时 2 列车的速度到站。假设火车到站的列数服从泊松分布，计算他平均等待列车的时间。

10. 某个投资产品的相关信息如下：

(1) 投资的利润率为当年证券回报指数的 75%，同时不低于 3% 的保证利润率。

(2) 年度证券回报指数服从期望为 8%，标准差为 16% 的正态分布。

(3) 对于一个期望值为 μ，标准差为 σ 的正态随机变量 X，给出期望值见表 4－4。

计算这个投资产品的年平均投资收益率。

表 4－4

$E[X \wedge 3\%]$		
	$\mu = 6\%$	$\mu = 8\%$
$\sigma = 12\%$	−0.43%	0.31%
$\sigma = 16\%$	−1.99%	−1.19%
$E[X \wedge 4\%]$		
	$\mu = 6\%$	$\mu = 8\%$
$\sigma = 12\%$	0.15%	0.95%
$\sigma = 16\%$	−1.43%	−0.58%

11. 一个四口之家，每人每年看病的次数服从均值为 1.5 的几何分布。每年每个家庭成员看病的次数相互独立。这个家庭购买了一份保险，从每人的第 4 次看病起，这份保险每次可以支付 100 元。计算这个家庭每年得到的平均赔付额。

12. 一个风险标的损失额服从均值为 3 的泊松分布。一份保单为这个风险提供保险保障，约定免赔额为 2；另一份保单的赔付比例为 α。假设这两份保单的平均成本相同，计算 α。

13. 某保险公司的公务用车保单条款如下：

(1) 每年的免赔额为 1 000 元；

(2) 被保险人赔付（1 000，6 000）区间内维修费用的 20%；

(3) 被保险人全额支付 6 000 元以上的维修费用，直到其总支付额达到 10 000 元；

(4) 被保险人每年再支付剩余维修费用的 10%。

假设汽车维修费用服从参数为 $\theta = 5\ 000$，$\alpha = 2$ 的帕累托分布。计算每年保险赔偿的期望值。

14. 随机变量 N 服从混合分布：

(1) 有 p 的可能性，N 服从 $q = 0.5, m = 2$ 的二项分布。

(2) 有 $1-p$ 的可能性，N 服从 $q = 0.5, m = 4$ 的二项分布。

求 $P(N = 2)$ 的表达形式。

15. 设某责任险实际损失额 X 的分布密度函数为 $f(x) = \frac{1}{100}\left(1 - \frac{x}{200}\right)$，$0 \leqslant x \leqslant 200$，保单约定如果实际损失额高于 50（百元），保险公司将赔偿损失额高于 50（百元）部分的 80%，同时还规定了保单限额为 150（百元），求理赔额 Y 的期望？

16. 设 Λ 是一个随机变量，服从均值为 1 的指数分布。已知给定 $\Lambda = \lambda$ 时，理赔次数 N 服从参数为 λ 的泊松分布，计算 $P(N = 0)$。

17. 对于某险种的理赔次数 N，有 $\frac{p_k}{p_{k-1}} = -\frac{1}{3} + \frac{4}{k}$，$k = 1, 2, 3, \cdots$，其中 $p_k = P(N = k)$，计算 $E(N)$ 和 $Var(N)$。

18. 设某险种的实际损失额有几种可能：25、50、75、100、200、500，发生的概率分别为 0.2、0.3、0.2、0.15、0.1、0.05，假设损失次数服从参数为 $r = 10$、$\beta = 0.3$ 的奇异负二项分布，免赔额为 50，求理赔次数的分布。

19. 健康保险中，医生对于控制病人的医疗费用起到关键作用。某保险公司为了控制医疗保险的赔付额，打算对医生实行一项奖励计划，如果每张保单的理赔额 Y 小于 400 元，医生就可以得到奖金 $c(400 - Y)$ 元。经验数据表明，理赔额 Y 服从参数 $\alpha = 2$、$\theta = 300$ 的帕累托分布，如果保险公司希望把支付给医生的平均奖金额控制在 100 元，c 应该是多少？

20. X 是一个属于（a，b，0）分布类的离散随机变量。已知：(1) $P(X = 0) = P(X = 1) = 0.25$；(2) $P(X = 2) = 0.1875$。计算 $P(X = 3)$。

21. 对于参数为 (α, θ) 的帕累托随机变量 X，证明：在均值 $E(X)$ 保持不变的条件下，令 $\alpha \to \infty$，则帕累托分布收敛到指数分布。

第五章　短期个体风险模型

学习目标

- □ 了解短期个体模型的研究对象以及内容
- □ 了解个体保单理赔分布的特点
- □ 熟悉研究保单组合理赔分布的三种方法：卷积方法、矩母函数方法以及近似方法
- □ 掌握用正态分布近似总理赔模型的方法

§5.1　引　　言

假设保险人在某个时间段（比如一个会计年度）内售出了 n 张保单，若被保险人在此期间发生了保险合同规定的保险事故，则保单持有者可以根据保险合同中承保的责任向保险人索赔，保险人则按合同的承诺进行赔付。对每份保单而言，在这个时间段内发生索赔与否不确定，赔付金额大小也不确定，要视损失程度与保险条款而定。假定第 i 张保单的理赔额为 X_i，则 X_i为非负随机变量，进而，保险人在这个时间段内的理赔或赔付总量为：

$$S = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i \tag{5.1.1}$$

一般情况下，要获得 S 的准确分布十分困难，因此我们首先在一些特殊的假设下讨论 S 的分布及其性质。假设如下：

假设 1：每张保单是否发生理赔以及理赔额的大小是相互独立的，即 X_1，X_2，…，X_n是独立的随机变量序列；

假设 2：每张保单至多发生一次理赔。若用随机变量 I 表示每张保单发生理赔的次数，则 I 的取值为 0 或 1，即 I 服从 0 - 1 分布或贝努里分布，记做：

$$I \sim \begin{pmatrix} 0 & 1 \\ 1-q & q \end{pmatrix}$$

其中，q 表示发生理赔的概率，它的确定要视具体问题而定，比如在寿险中往往由生命表确定。

假设 3：保单总数 n 是事先确定的正整数。

我们称满足以上 3 个假设的式（5.1.1）为**个体风险模型**，这个模型以

研究随机变量 S 的分布情况为主要内容。

由于理论不可能只用一个模型来概括全部现实和适应各种类型的实际问题，因此以上假设都是对实际情况的简化和理想化。独立性假设 1 是应用大数法则基本原理的前提，可以看做是最基本的假设，尽管它并不概括所有的情况，如水灾保险和传染病保险等等；假设 2 在某些寿险和非寿险情形可能具有一定的代表性，但显然不是每种保单都只允许索赔一次，例如汽车保险和健康保险就可能会发生多次索赔；假设 3 似乎与实际情况相差甚远，几乎纯粹是为了数学处理上的方便，但是也可以把它们看做是研究实际问题的基础，实际的保单组合可能包含一系列保单子类，而某些保单子类可能近似符合假设 3 的情况。在这个模型中，由于保单数事先确定，因此也称之为封闭模型。

§5.2　个体保单的理赔分布

【例 5－1】　考虑某种一年期的寿险保单组合，保单规定若投保人在一年之内意外身亡，保险人将赔付 b 元，若无意外则不予赔付。假设被保险人在一年内意外死亡的概率为 q，求每份保单理赔额的均值和方差？

解： 记 X_i为保险人对第 i 张保单的赔付额，则 $X_i \sim \begin{pmatrix} 0 & b \\ 1-q & q \end{pmatrix}$，若用 $F_i(x)$ 表示 X_i的分布函数，则有

$$F_i(x) = P\{X_i \leqslant x\} = \begin{cases} 0, & x<0 \\ 1-q, & 0 \leqslant x < b \\ 1, & x \geqslant b \end{cases}$$

记 I 为理赔次数，则 I 服从贝努里分布，即

$$I \sim \begin{pmatrix} 0 & 1 \\ 1-q & q \end{pmatrix} \tag{5.2.1}$$

因此，X_i又可以表示为 $X_i = Ib$，易知 $E[I]=q$，$Var[I]=q(1-q)$；对于 X_i，有

$$E[X_i]=bq, \qquad Var[X_i]=b^2q(1-q) \tag{5.2.2}$$

■

【例 5－2】　对例 5－1 中的赔付规则作如下修改：若保单持有人在一年保险期内因"意外"身故，赔付额为 10 万元；若因"非意外"而身故，则赔付 5 万元；若未发生身故事件则合同自然终止。根据历史数据记录，"意外"和"非意外"身故的概率分别为 0.0005 和 0.0020。这里假定"意外"身故和一般身故的发生是独立的，试重新讨论第 i 张保单理赔额的概率分布。

解： 仍用 I 表示理赔次数，$I=l$ 表示有死亡事故发生需要赔付，$I=0$

则表示无事故发生不需要赔付；若用 B 表示需要赔付的数额，B 已不再是常数，而是与 I 有关的随机变量，依题意有：

$$P\{I=1,B=100\ 000\}=0.0005,\ P\{I=1,B=50\ 000\}=0.0020$$

而且，

$$q=P\{I=1\}=P\{I=1,B=100\ 000\}+P\{I=1,B=50\ 000\}=0.0025$$

$$1-q=1-P\ \{I=1\}\ =0.9975$$

因此，仍记 $X_i=IB$，其中 B 的条件分布为：

$$P\{B=50\ 000\mid I=1\}=\frac{P\{I=1,B=50\ 000\}}{P\{I=1\}}=\frac{0.0020}{0.0025}=0.8$$

$$P\{B=100\ 000\mid I=1\}=\frac{P\{I=1,B=100\ 000\}}{P\{I=1\}}=\frac{0.0005}{0.0025}=0.2$$

并且有

$$\begin{aligned}E[X_i]&=P\{I=0\}E[X_i\mid I=0]+P\{I=1\}E[X_i\mid I=1]\\&=(1-q)\times 0+qE[B\mid I=1]\\&=q(50\ 000P\{B=50\ 000\mid I=1\}+100\ 000P\{B=100\ 000\mid I=1\})\\&=150\end{aligned}$$

$$\begin{aligned}Var[X_i]&=E(X_i^2)-[E(X_i)]^2\\&=P\{I=0\}E(X_i^2\mid I=0)+P\{I=1\}E(X_i^2\mid I=1)-(E[X_i])^2\\&=qE(B^2\mid I=1)-(E[X_i])^2\\&=0.0025\times E(B^2\mid I=1)-150^2\\&=0.0025(50\ 000^2\times 0.8+100\ 000^2\times 0.2)-150^2\\&=9\ 977\ 500\end{aligned}$$

一般地，若随机变量 X 可表示为两个随机变量 I 和 B 的乘积：$X=IB$，则由

$$E(X)=E[E(X\mid I)]$$

$$Var(X)=Var[E(X\mid I)]+E[Var(X\mid I)]$$

这里 I 为贝努里变量，I 与 B 独立，记 $q=P\ (I=1)$，因此有

$$E(X)=qE(B) \tag{5.2.3}$$

$$Var(X)=q(1-q)E^2(B)+qVar(B) \tag{5.2.4}$$

因此，可以利用式（5.2.3）和（5.2.4）重新计算例 5－2。

已知：$q=P(I=1)=0.0052$，$P(B=5\ 000)=4P(B=100\ 000)$，若定义 50 000 为一个货币单位，则有 $P(B=1)=\frac{4}{5}=1-P(B=2)$，因此有

$$E(B)=1\times\frac{4}{5}+2\times\frac{1}{5}=\frac{6}{5},\ E(B^2)=1^2\times\frac{4}{5}+2^2\times\frac{1}{5}=\frac{8}{5}$$

将这些中间结果代入式（5.2.3）和（5.2.4）可以得到与例 5－2 同样的结果。 ■

【例 5－3】　设有某种汽车车辆险保单，赔付规则设定免赔额为 250 元，最高赔付额为 2 000 元，还假定在保险期内至多有一次索赔，且 $P\{I=1\}=0.15$。由于损失超过 2 250 元后最多赔付 2 000 元，因此对索赔额 B 假定：$P\{B=2\,000 \mid I=1\}=0.1$，而在 2 000 元以下部分的概率分布函数为：

$$P\{B\leqslant x \mid I=1\}=\begin{cases}0, & x<0\\ 0.9\left[1-\left(1-\dfrac{x}{2\,000}\right)^2\right], & 0\leqslant x<2\,000\\ 1, & x\geqslant 2\,000\end{cases}$$

试讨论 X 的概率分布。

解：记理赔额 X 的分布函数为 $F(x)$，则

$$\begin{aligned}F(x)&=P\{X\leqslant x\}\\&=P\{IB\leqslant x\}\\&=P\{IB\leqslant x \mid I=0\}P\{I=0\}+P\{IB\leqslant x \mid I=1\}P\{I=1\}\\&=\begin{cases}0\times0.85+0\times0.15, & x<0\\ 1\times0.85+0.9\left[1-\left(1-\dfrac{x}{2\,000}\right)^2\right]\times0.15, & 0\leqslant x<2\,000\\ 1\times0.85+1\times0.15, & x\geqslant 2\,000\end{cases}\\&=\begin{cases}0, & x<0\\ 0.985-0.135\left(1-\dfrac{x}{2\,000}\right)^2, & 0\leqslant x<2\,000\\ 1, & x\geqslant 2\,000\end{cases}\end{aligned}$$

这是一个不连续的函数，又称"混合分布函数"，可以看做由离散和连续的概率分布混合而成，它在区间 $0<x<2\,000$ 上的概率密度为：

$$f(x)=0.000135\left(1-\frac{x}{2\,000}\right),\quad 0<x<2\,000$$

而且 $P(X=0)=0.85$，$P(X=2\,000)=0.015$，所以 X 的 k 阶原点矩为：

$$E(X^k)=\int_0^{2\,000}x^kf(x)dx+2\,000^k\times P(X=2\,000)$$

由此可以算出 X 的均值和方差分别为：

$$E(X)=0.000135\int_0^{2\,000}x\left(1-\frac{x}{2\,000}\right)dx+2\,000\times0.015=120$$

$$\begin{aligned}Var(X)&=E(X^2)-[E(X)]^2\\&=0.000135\int_0^{2\,000}x^2\left(1-\frac{x}{2\,000}\right)dx+2\,000^2\times0.015-120^2\\&=135\,600\end{aligned}$$

■

§5.3　总理赔额的分布——卷积法

在前面讨论个体保单理赔分布的基础上，我们开始考虑保单组合的理赔分

布，即独立随机变量和 $S=\sum_{i=1}^{n}X_i$ 的分布，本节讨论第一种方法——卷积方法。

5.3.1 两项的卷积

【例 5－4】 相互独立的两个离散随机变量 X_1 和 X_2，其概率分布分别为：

$$X_1\sim\begin{pmatrix}0 & 1 & 2 & \cdots & m\\ p_0 & p_1 & p_2 & \cdots & p_m\end{pmatrix},\ X_2\sim\begin{pmatrix}0 & 1 & 2 & \cdots & n\\ q_0 & q_1 & q_2 & \cdots & q_n\end{pmatrix}$$

试求 $S=X_1+X_2$ 的分布。

解： $S=X_1+X_2$ 的可能取值为 $\{0,1,\cdots,m+n\}$；对应的概率为：

$$\begin{aligned}P\{S=k\}&=P\{X_1+X_2=k\}\\&=P\{(X_1=0,X_2=k)\cup(X_1=1,X_2=k-1)\cup\cdots(X_1=k,X_2=0)\}\\&=\sum_{i=0}^{k}P\{X_1=i,X_2=k-i\}=\sum_{i=0}^{k}P\{X_1=i\}P\{X_2=k-i\}\\&=\sum_{i=0}^{k}p_iq_{k-i}\end{aligned}$$

或者

$$\begin{aligned}P\{S=k\}&=P\{X_1+X_2=k\}\\&=P\{(X_1=k,X_2=0)\cup(X_1=k-1,X_2=1)\cup\cdots(X_1=0,X_2=k)\}\\&=\sum_{j=0}^{k}p_{k-j}q_j\end{aligned}$$

■

【例 5－5】 相互独立的两个非负连续索赔额随机变量 X_1 和 X_2，设联合概率密度和边际密度分别为 $f(x,y)$、$f_1(x)$ 和 $f_2(y)$，试求 $S=X_1+X_2$ 的概率密度。

解： 由独立性知 $f(x,y)=f_1(x)f_2(y)$，设 S 的分布函数和密度分别为 $F_S(z)$ 和 $f_S(z)$，由定义有：

$$\begin{aligned}F_S(z)&=P\{X_1+X_2\leqslant z\}\\&=\iint\limits_{x+y\leqslant z}f(x,y)dxdy=\iint\limits_{x+y\leqslant z}f_1(x)f_2(y)dxdy\\&=\int_0^z f_1(x)\left[\int_0^{z-x}f_2(y)dy\right]dx=\int_0^z f_1(x)\left[\int_x^z f_2(y-x)dy\right]dx\\&=\int_0^z\left[\int_0^y f_1(x)f_2(y-x)dx\right]dy\end{aligned}$$

所以，$S=X_1+X_2$ 的分布密度为：

$$f_S(z)=\int_0^z f_1(x)f_2(z-x)dx \qquad (5.3.1)$$

或者

$$f_S(z)=\int_0^z f_1(z-y)f_2(y)dy \qquad (5.3.2)$$

式（5.3.2）称为**卷积公式**，记为 $f_S(z)=f_1*f_2(z)$。■

【例 5－6】 设 X 在$[0,100]$上均匀分布，Y 在$[0,150]$上均匀分布，X 与 Y 相互独立，令 $S=X+Y$，试计算 $f_S(175)$。

解： 由式（5.3.1），$f_S(175)=\int_0^{175} f_1(x)f_2(175-x)dx$。又由 X 和 Y 分布的定义，有

$$f_1(x)=\begin{cases}0, & x<0\\ \frac{1}{100}, & 0\leqslant x\leqslant 100\\ 0, & x>100\end{cases} \qquad f_2(y)=\begin{cases}0, & y<0\\ \frac{1}{150}, & 0\leqslant y\leqslant 150\\ 0, & y>150\end{cases}$$

因此有

$$\begin{aligned}f_S(175)&=\int_{25}^{100} f_1(x)f_2(175-x)dx\\&=\frac{1}{100}\times\frac{1}{150}\times(100-25)=\frac{1}{200}\end{aligned}$$

■

5.3.2　多项卷积

【例 5－7】 设 X_1，X_2，…，X_n为相互独立的 n 个离散随机变量（$n>2$），并设 X_i的概率函数为 $p_i(z)$，$i=1,2,\cdots,n$。试求多项和 $S=X_1+X_2+\cdots+X_n$的概率分布函数。

解： 利用卷积方法求多个独立随机变量和的分布，可以在两项卷积的基础上通过逐次迭代来实现，比如要求 $S=X_1+X_2+X_3$的分布，可以先对 X_1和 X_2进行卷积，然后再将 X_1+X_2和 X_3进行卷积运算，如此迭代下去。

对于 $1<k<n$，记 $X_1+X_2+\cdots+X_k$的概率函数为 $p^{*(k)}(z)$，则有

$$p^{*(k+1)}(z)=p^{*(k)}*p_{k+1}(z)$$

■

【例 5－8】 设随机变量 X_1，X_2和 X_3相互独立，且 $X_i\sim\begin{pmatrix}1 & 2 & 3\\ 0.5 & 0.4 & 0.1\end{pmatrix}$，试求 $S=X_1+X_2+X_3$的概率分布。

解： $X_1\sim p^{*1}(x)=p_1(x)=\begin{pmatrix}1 & 2 & 3\\ 0.5 & 0.4 & 0.1\end{pmatrix}$

$$X_1+X_2\sim p^{*2}(x)=p_1*p_1(x)=\begin{pmatrix}2 & 3 & 4 & 5 & 6\\ 0.25 & 0.4 & 0.26 & 0.08 & 0.01\end{pmatrix}$$

$$\begin{aligned}X_1+X_2+X_3&=(X_1+X_2)+X_3\sim p^{*3}(x)=p_1*p_1*p_1(x)\\&=\begin{pmatrix}1 & 2 & 3\\ 0.5 & 0.4 & 0.1\end{pmatrix}*\begin{pmatrix}2 & 3 & 4 & 5 & 6\\ 0.25 & 0.4 & 0.26 & 0.08 & 0.01\end{pmatrix}\end{aligned}$$

具体计算结果如表 5－1 所示。

表 5－1　　　　例 5－8 的计算过程

x	$p^{*0}(x)$	$p^{*1}(x)=p_1(x)$	$p^{*2}(x)$	$f(x)=p^{*3}(x)$	$F(x)$
1	—	0.5	—	—	—
2	—	0.4	0.25	—	—
3	—	0.1	0.40	0.125	0.125
4	—	—	0.26	0.300	0.425
5	—	—	0.08	0.315	0.740
6	—	—	0.01	0.184	0.924
7	—	—	—	0.063	0.987
8	—	—	—	0.012	0.999
9	—	—	—	0.001	1.0000

■

【例 5－9】 设随机变量 X_1，X_2和 X_3相互独立，其概率分布依次如下：

$$X_1 \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{pmatrix}, \quad X_2 \sim \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0.5 & 0.2 & 0.1 & 0.1 & 0.1 \end{pmatrix},$$

$$X_3 \sim \begin{pmatrix} 0 & 2 & 3 & 4 & 5 \\ 0.6 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}$$

试求 $S=X_1+X_2+X_3$的概率分布。

解： 由两项和卷积公式（5.3.1），有：

$$X_1+X_2 \sim \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0.2 & 0.23 & 0.2 & 0.16 & 0.11 & 0.06 & 0.03 & 0.01 \end{pmatrix}$$

运用迭代方法有：

$$X_1+X_2+X_3=(X_1+X_2)+X_3 \sim p^{*(3)}(z)=p^{*(2)} * p_3(z)$$

$$= \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 0.12 & 0.138 & 0.14 & 0.139 & 0.129 & 0.115 & 0.088 & 0.059 & 0.036 & 0.021 & 0.01 & 0.004 & 0.001 \end{pmatrix}$$

计算结果综合列入表 5－2。

表 5－2　　　　例 5－9 的计算过程

z	$p_1(z)$	$F_1(z)$	$p_2(z)$	$P^{*(2)}(z)$	$F^{*(2)}(z)$	$p_3(z)$	$p^{*(3)}(z)$	$F^{*(3)}(z)$
0	0.4	0.4	0.5	0.20	0.20	0.6	0.120	0.120
1	0.3	0.7	0.2	0.23	0.43	0.0	0.138	0.258
2	0.2	0.9	0.1	0.20	0.63	0.1	0.140	0.398
3	0.1	1.0	0.1	0.16	0.79	0.1	0.139	0.537
4	0.0	1.0	0.1	0.11	0.90	0.1	0.129	0.666

续表

z	$p_1(z)$	$F_1(z)$	$p_2(z)$	$p^{*(2)}(z)$	$F^{*(2)}(z)$	$p_3(z)$	$p^{*(3)}(z)$	$F^{*(3)}(z)$
5	0.0	1.0	0.0	0.06	0.96	0.1	0.115	0.781
6	0.0	1.0	0.0	0.03	0.99	0.0	0.088	0.868
7	0.0	1.0	0.0	0.01	1.00	0.0	0.059	0.928
8	0.0	1.0	0.0	0.00	1.00	0.0	0.036	0.964
9	0.0	1.0	0.0	0.00	1.00	0.0	0.021	0.985
10	0.0	1.0	0.0	0.00	1.00	0.0	0.010	0.995
11	0.0	1.0	0.0	0.00	1.00	0.0	0.004	0.999
12	0.0	1.0	0.0	0.00	1.00	0.0	0.001	1.00

■

同样，也可以计算多个连续的独立随机变量和的概率分布函数和密度函数。

【例 5－10】 设 X_1，X_2，…，X_n为相互独立且同分布的 n 个连续随机变量（$n>2$），试求多项和 $S=X_1+X_2+\cdots+X_n$的概率分布函数和概率密度。

解： 设 X_i的分布函数为 $F(x)$，分布密度为$f(x)$，$i=1$，2，…，n，则 S 的概率分布函数 $F_S(x)$和概率密度$f_S(x)$分别为：

$$F_S(x)=F^{*n}(x)=F^{*(n-1)}*F(x)$$

$$f_S(x)=f^{*n}(x)=f^{*(n-1)}*f(x)$$

其中：$F^{*1}(z)=F(z)$，$f^{*1}(z)=f(z)$

因此，$F^{*2}(z)=F*F(z)=P\{X_1+X_2\leqslant z\}$，$f^{*2}(z)=f*f(z)$，依此类推。■

§5.4　总理赔额的分布——矩母函数法

卷积方法计算理赔变量的精确分布往往需要进行大量的计算。而对于取值非负的随机变量 X，由矩母函数的定义 4.1 可知概率分布函数 $F(x)$，$x\geqslant 0$ 与函数 $M_X(t)$，$t\geqslant 0$ 是一一对应的，互相决定。

由矩母函数的性质 3 我们还可以知道，独立随机变量和的矩母函数与各随机变量矩母函数之间的关系，即：对于独立随机变量和 $S=X_1+X_2+\cdots+X_n$，有

$$M_S(t)=M_{X_1}(t)\cdots M_{X_n}(t) \tag{5.4.1}$$

若 X_1，X_2，…，X_n 独立同分布，设其共同的矩母函数为 $M_X(t)$，则

$$M_S(t)=[M_X(t)]^n,\qquad t\geqslant 0 \tag{5.4.2}$$

根据矩母函数的性质（5.4.1）和（5.4.2），往往可以比较容易地获

得保单组合理赔分布的矩母函数，然后利用矩母函数与分布一一对应的性质获得总理赔的分布函数。例如，第四章的定理 4－2 就说明了独立的泊松随机变量之和仍然服从泊松分布，并且参数为相应的各参数之和。事实上，服从负二项分布的随机变量也具有类似的性质。

定理 5－1 设 X_1，X_2，…，X_n为相互独立的随机变量，X_i服从参数为（r_i，p）的负二项分布，则 $S=X_1+X_2+\cdots+X_n$服从参数为（$\sum_{i=1}^{n} r_i$，p）的负二项分布。

证明： 由第四章对负二项分布的介绍，可知 X_i的概率母函数为：

$$P_{X_i}(t)=\left(\frac{p}{1-qt}\right)^{r_i},\qquad 0<t<\frac{1}{q}$$

因此矩母函数为：

$$M_{X_i}(t)=P_{X_i}(e^t)=\left(\frac{p}{1-qe^t}\right)^{r_i},\qquad 0<t<-\ln q$$

所以根据式（5.4.1），S 的矩母函数为：

$$M_S(t)=\prod_{i=1}^{n}M_{X_i}(t)=\prod_{i=1}^{n}\left(\frac{p}{1-qe^t}\right)^{r_i}=\left(\frac{p}{1-qe^t}\right)^{\sum_{i=1}^{n}r_i},t<-\ln q$$

再次对照负二项分布的矩母函数可知，S 服从参数为（$\sum_{i=1}^{n} r_i$，p）的负二项分布。

同理可证 r 个独立同分布的参数为 p 的几何分布之和服从参数为(r,p)的负二项分布。 ■

对于连续随机变量，我们在第四章也有如下结果：如果 X_1，…，X_n是独立的且服从参数为(α_i,θ)的伽玛分布随机变量，则 $Y=\sum_{i=1}^{n}X_i$ 服从参数为（$\sum_{i=1}^{n}\alpha_i$，θ）的伽玛分布。特别地，若 X_1，…，X_n是独立且同服从参数为 θ 的指数分布，则 $Y=\sum_{i=1}^{n}X_i$ 服从参数为（n，θ）的伽玛分布，这说明在固定尺度参数 θ 的条件下，伽玛分布具有可加性，事实上，正态分布也具有可加性。

定理 5－2 设 X_1，$X_2,\cdots,X_n$为相互独立的随机变量，$X_i\sim N$（μ_i，σ_i^2），$i=1$，$2,\cdots,n$，则 $S=X_1+X_2+\cdots+X_n\sim N\left(\sum_{i=1}^{n}\mu_i,\sum_{i=1}^{n}\sigma_i^2\right)$。

证明： 由第四章正态分布的矩母函数以及式（5.4.2）知 S 的矩母函数为：

$$M_S(t)=\prod_{i=1}^{n}M_{X_i}(t)=\prod_{i=1}^{n}e^{\mu_i t+\frac{1}{2}\sigma_i^2t^2}=e^{\sum_{i=1}^{n}\mu_i t+\frac{1}{2}\sum_{i=1}^{n}\sigma_i^2t^2}$$

对照正态分布的矩母函数可知，S 服从参数为$\left(\sum_{i=1}^{n}\mu_i,\sum_{i=1}^{n}\sigma_i^2\right)$的正态分布。

■

定理 5－1 和 5－2 表明这些分布族对于独立和的运算是封闭的，即分布族中的独立随机变量和的分布仍然属于该分布族。如果将这个性质用于个体风险模型，也就是个体理赔的分布与总理赔的分布属于同一个分布族。

§5.5　总理赔额分布的正态近似

卷积方法和矩母函数法都可以获得独立随机变量和的精确分布，但对于保单数较多的保单组合来说，更实用的方法则是考虑组合理赔量的近似分布。

从概率论的中心极限定理知，在一定的条件下，当变量的个数趋向无穷时，独立随机变量和的分布将趋于正态分布。中心极限定理的具体表述为：

中心极限定理　设 X_1，$X_2,\cdots,X_n$为独立同分布的随机变量序列，并且 $E(X_n)=\mu<\infty$，$Var(X_n)=\sigma^2<\infty$。则 $S=X_1+X_2+\cdots+X_n$的分布将趋向于正态分布，即 $\zeta_n=\dfrac{S-E(S)}{\sqrt{Var(S)}}=\dfrac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)$ 满足

$$\lim_{n\to\infty}P(\zeta_n\leqslant x)=\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt \tag{5.5.1}$$

中心极限定理中的同分布条件并不是必要条件，一般地，设 X_1，X_2，$\cdots$，X_k，$\cdots$ 为独立随机变量系列，设 X_k的分布函数为 $F_k(x)$，且 $E(X_k)=\mu_k$，$Var(X_k)=\sigma_k^2$，记 $B_n^2=\sum_{k=1}^{n}\sigma_k^2$，$\zeta_n=\sum_{k=1}^{n}\dfrac{X_k-\mu_k}{B_n}$，则式（5.5.1）成立的更为一般的充分条件是如下的林德贝格条件：

$$\lim_{n\to\infty}\frac{1}{B_n^2}\sum_{k=1}^{n}\int_{|x-\mu_k|>\varepsilon B_n}(x-\mu_k)^2dF_k(x)=0 \qquad 对任意\ \varepsilon>0 \tag{5.5.2}$$

因为风险理论主要是应用中心极限定理，所以，这里我们略去了有关的证明，请读者自行参阅有关的文献。

有了上面的定理，我们可以用来计算保单数很多的保单组合的总理赔分布。基本的计算步骤是：

第一，利用个体理赔的分布计算总理赔 S 的均值 $E(S)=\sum_{k=1}^{n}E(X_k)$ 和方差 $Var(S)=\sum_{k=1}^{n}Var(X_k)$；

第二，对 S 的分布计算进行标准化处理：

$$P(S\leqslant s)=P\left(\frac{S-E(S)}{\sqrt{Var(S)}}\leqslant\frac{s-E(S)}{\sqrt{Var(S)}}\right)$$

第三，利用中心极限定理近似计算：

$$P(S \leqslant s) \approx \Phi\left(\frac{s-E(S)}{\sqrt{Var(S)}}\right)$$

因为进行理赔分析的目的之一是确定保费，所以通常会考虑 s 为总理赔 S 的均值的某个倍数，例如，$s=(1+\theta)E(S)$，这时称 $\theta E(S)$ 为该保单组合的安全附加保费，称 θ 为相对附加安全系数（或安全附加保费率）。

【例 5－11】 考虑由 10 万张同类医疗保险保单构成的组合。假设承保的损失相互独立，按照保单的规定，保险人将负责赔付所发生损失的 80%。设每张保单在保险期内的损失均服从如表 5－3 所示的分布。若要求收取的保费总额低于理赔总额的概率不超过 5%，试确定该保单组合最低的安全附加保费。

表 5－3　　　　个体保单的损失分布

X	0	50	200	500	1 000	10 000
$P\{X=x\}$	0.30	0.10	0.10	0.20	0.20	0.10

解： 经计算，有

$$E(X)=1\ 325, \qquad Var(X)=8\ 498\ 625$$

设总损失变量为：

$$L=\sum_{i=1}^{100\ 000} X_i$$

则 $$E(L)=1\ 325\times 100\ 000, \qquad Var(L)=8\ 498\ 625\times 100\ 000$$

理赔总额变量为：

$$S=0.8L=0.8\sum_{i=1}^{100\ 000} X_i$$

则 $$E(S)=0.8\times 1\ 325\times 100\ 000=1\ 060\times 100\ 000$$

$$Var(S)=0.8^2\times 8\ 498\ 625\times 100\ 000=5\ 439\ 120\times 100\ 000$$

又设保费总额为：

$$G=(1+\theta)E(S)$$

按题意要求，G 应该满足：

$$P\{S\leqslant(1+\theta)E(S)\}\geqslant 95\%$$

也就是说，

$$s=(1+\theta)E(S) \Rightarrow \Phi\left(\frac{s-E(S)}{\sqrt{Var(S)}}\right)=\Phi\left(\frac{\theta E(S)}{\sqrt{Var(S)}}\right)$$

若 $\Phi\left(\frac{\theta E(S)}{\sqrt{Var(S)}}\right)=0.95$，则有 $\frac{\theta E(S)}{\sqrt{Var(S)}}=1.645$，则该保单组合的安全附加保费为：

$$\theta E(S)=1.645\times\sqrt{5\ 439\ 120\times 100\ 000}=1\ 213\ 193.9$$

相对附加安全系数 θ 为：$\theta=\dfrac{1\ 213\ 193.9}{1\ 060\times 100\ 000}\approx 0.011$。 ■

【例 5.12】 设某保险公司共售出某种汽车保单 2 500 张，保单持有者被分作两类，情况如表 5－4 所示。其中，理赔额 B_k 服从参数为（λ，L）的截断指数分布，分布密度函数为：

$$f(x)=\begin{cases}0, & x\leqslant 0\\ \lambda e^{-\lambda x}, & 0<x<L\\ e^{-\lambda L}, & x=L\end{cases} \tag{5.5.3}$$

表 5－4　保单持有者风险状况

类型 k	人数 n_k	理赔概率 q_k	理赔额 B_k 的分布参数	
			λ	L
1	500	0.10	1	2.5
2	2 000	0.05	2	5.0

若要求收取的保费总额低于理赔总额的概率不超过 5%，试计算安全附加保费率 θ。

解： 与例 5－11 类似，用 S 表示理赔总量，要计算概率不等式 $P\{S\leqslant(1+\theta)E(S)\}\geqslant 95\%$ 中的 θ。

根据正态近似，同样有

$$\theta=1.645\times\frac{\sqrt{Var(S)}}{E[S]} \tag{5.5.4}$$

因此，关键是计算 $E(S)$ 和 $Var(S)$。记 $S=\sum_{i=1}^{2\,500}X_i=\sum_{i=1}^{2\,500}I_iB_i=\sum_{i=1}^{500}I_iB_i+\sum_{j=501}^{2\,500}I_jB_j$，其中 B_i 和 B_j 分别对应两种不同情况下的理赔额。对于参数为（λ，L）的截断指数分布 B，有

$$E(B)=\int_0^L x\lambda e^{-\lambda x}dx+Le^{-\lambda L}=\frac{1-e^{-\lambda L}}{\lambda},$$

$$\begin{aligned}Var(B)&=E[B^2]-(E[B])^2\\&=\int_0^L x^2\lambda e^{-\lambda x}dx+L^2e^{-\lambda L}-\left(\frac{1-e^{-\lambda L}}{\lambda}\right)^2\\&=\frac{1-2\lambda Le^{-\lambda L}-e^{-2\lambda L}}{\lambda^2}\end{aligned}$$

再分别用（λ，L）＝（1，2.5）和（λ，L）＝（2，5）代入上式得：

$$E(B_i)=0.9179,Var(B_i)=0.5828,i=1,2,\cdots,500$$

$$E(B_j)=0.5,Var(B_j)=0.2498,j=501,502,\cdots,2\ 500$$

从而有：

（1）当 $i=1,\ 2,\ \cdots,\ 500$ 时，

$$E(X_i)=E(I_iB_i)=E(E[B_i\mid I_i])=q_1E(B_i\mid I_i=1)=0.09179$$

$$\begin{aligned}Var(X_i)&=Var(I_iB_i)=Var(E[B_i\mid I_i])+E[Var(B_i\mid I_i=1)]\\&=(E[B\mid I_i=1])^2q_1(1-q_1)+q_1Var(B_i\mid I_i=1)=0.13411\end{aligned}$$

（2）类似地，当 $j=501,\ 502,\ \cdots,\ 2\ 500$ 时，

$$E(X_j)=0.025, Var(X_j)=0.02436$$

从而有：

$$E(S)=500\times 0.09179+2\ 000\times 0.025=95.90$$

$$Var(S)=500\times 0.13411+2\ 000\times 0.02436=115.78$$

因此，$\theta=1.645\times\frac{\sqrt{Var(S)}}{E[S]}=1.645\times\frac{\sqrt{115.78}}{95.89}=0.1846$

■

【例 5－13】 某团体医疗保险业务的理赔数据如表 5－5 所示（以年为时间单位）。

试计算理赔总额超过 18 万元的概率（用正态近似）。

表 5－5 个体理赔额经验数据

保单类型	投保人数	个体理赔额（单位：元）	
		均值	标准差
个人	786	76	42
家庭	592	187	77

解：用 S 表示理赔总额，则

$$E(S)=786\times 76+592\times 187=170\ 440$$

$$Var(S)=786\times 42^2+592\times 77^2=2\ 212.8^2$$

因此，

$$P(S>180\ 000)=P\left(\frac{S-170\ 440}{\sqrt{(2\ 212.8)^2}}>\frac{180\ 000-170\ 440}{\sqrt{(2\ 212.8)^2}}\right)=1-\Phi(4.32)$$
$$=7.80\times 10^{-6}$$

【例 5－14】 某公司为其 14 名员工购买了“一年期团体定期寿险”，每位员工的给付额根据工资水平确定，死亡率以中国人寿保险业经验生命表 CL00－03 为准（见表 5－6）。试按照精确计算和正态近似的方法计算理赔总额超过 5 万元的概率，比较两个结果的差异并分析原因。

表 5－6 承保员工的相关数据

雇员 j	年龄	性别	给付额 b_j（单位：万元）	死亡率 q_j	雇员 j	年龄	性别	给付额 b_j（单位：万元）	死亡率 q_j
1	20	F	1	.000283	8	27	M	8	.000795
2	21	M	2	.000661	9	28	M	9	.000815
3	22	M	3	.000692	10	29	M	10	.000842
4	23	F	4	.000328	11	30	F	10	.000406
5	24	M	5	.000738	12	31	M	15	.000932
6	25	F	6	.000347	13	32	F	15	.000465
7	26	M	7	.000779	14	33	M	20	.001055
合计			28					87	

解：首先考虑 $P(S>5)$ 的精确值。记 S 的概率分布函数为 $p_S(x)$，由于 $S=\sum_{i=1}^{14}X_i$，所以有 $p_S(x)=p_{X_1}*p_{X_2}*\cdots*p_{X_{14}}(x)$，其中，

$$p_{X_i}(x)=\begin{cases}p_j, & x=0\\ q_j, & x=b_j\end{cases},\qquad j=1,2,\cdots,14$$

若记 $S_j=S_{j-1}+X_j$，$j=2,3,\cdots,14$，则 S 的概率函数可以通过迭代的卷积计算得到。

从 $S_1=X_1$开始，得：

$$p_{S_j}(x)=\begin{cases}p_jp_{S_{j-1}}(x), & x<b_j\\ p_jp_{S_{j-1}}(x)+p_{S_{j-1}}(x-b_j)q_j, & x\geqslant b_j\end{cases},\quad j=2,3,\cdots,14$$

关于 S_1和 S_2的详细结果如下：

$p_{S_1}(0)=0.999717$，　　$p_{S_1}(1)=0.000283$

$p_{S_2}(0)=p_2\times f_{S_1}(0)=0.999056$，　　$p_{S_2}(1)=p_2\times p_{S_1}(1)=0.000283$

$p_{S_2}(2)=q_2\times p_{S_1}(0)=0.00661$，　　$p_{S_2}(3)=q_2\times p_{S_1}(1)=1.87\times10^{-7}$

依此继续，可得 $p_S(x)$的值，表 5－7 给出了 $F_S(x)$的值。

表 5－7　　通过迭代方法计算得到的 $F_S(x)$的精确值

x	0	1	2	3	4	5	6	7
$F_S(x)$	0.9909	0.9909	0.9916	0.9922	0.9926	0.9933	0.9936	0.9944
x	8	9	10	11	12	13	14	15
$F_S(x)$	0.9952	0.9960	0.9973	0.9973	0.9973	0.9973	0.9973	0.9989

因此有：

$$F_S(5)=0.9933, 1-F_S(5)=0.0067$$

其次考虑 $P(S>5)$的近似值，因为

$$E(S)=\sum_{j=1}^{14}q_jb_j=0.084$$

$$Var(S)=\sum_{j=1}^{14}q_jb_j^2-\sum_{j=1}^{14}q_j^2b_j^2=\sum_{j=1}^{14}q_j(1-q_j)b_j^2=1.06061=(1.02989)^2$$

按照正态近似，

$$\begin{aligned}P(S>5)&=1-P(S\leqslant5)\\&=1-P\left(\frac{S-0.08445}{1.02986}\leqslant\frac{5-0.08445}{1.02986}\right)\\&=1-\Phi(4.77303)=9.07\times10^{-7}\end{aligned}$$

比较两种方法的结果，显然，0.0067 比 9.6×10^{-7}大很多。

产生这一偏差的主要原因是，S 的分布不是像正态分布那样的对称分布，而是有偏分布，经计算其偏度为 $\mu_3(\mu_2)^{-\frac{3}{2}}=12.55$，非常大，因此用

正态分布来近似并不恰当。

一般来说，只有当样本量 n 非常大时，才适合用正态分布来近似其总分布，这个题目的样本量只有 14，显然过小，因此使用正态分布近似是不合适的。 ■

习　题

1. 假设 X 是连续扔 5 次硬币后"国徽"面朝上的次数，然后再一起扔 X 个骰子。设 Y 是 X 个骰子的数目总和，试求 Y 的均值和方差。

2. 某建筑物的价值为 a，在一定时期内发生火灾的概率为 0.02。火灾发生后，建筑物的损失额服从 0 到 a 的均匀分布。试计算在该时期内损失额的均值和方差。

3. X_1和 X_2是两个独立随机变量，取值都是非负整数。$S = X_1 + X_2$。已知表 5－8，计算 $f_S(1)$。

表 5－8

x	$F_1(x)$	$F_2(x)$
0	0.50	0.20
1	0.80	0.42
2	0.90	0.78
3	1.00	0.92

4. X_i，$i=1$，2，3 是独立同分布的随机变量，共同的分布函数为：

$$F(x) = \begin{cases} 0, & x<0 \\ x, & 0 \leqslant x < 1 \\ 1, & x \geqslant 1 \end{cases}$$

记 $S = X_1 + X_2 + X_3$。要求：

(1) 证明 S 的分布函数为：

$$F_S(x) = \begin{cases} 0, & x<0 \\ \dfrac{x}{6}, & 0 \leqslant x < 1 \\ [x^3 - 3(x-1)^3]/6, & 1 \leqslant x < 2 \\ [x^3 - 3(x-1)^3 + 3(x-2)^3]/6, & 2 \leqslant x < 3 \\ 1, & x \geqslant 3 \end{cases}$$

(2) 计算 $E(S)$和 $Var(S)$；

(3) 用 (1) 中的分布函数计算下列概率：(*i*) $P(S \leqslant 0.5)$；(*ii*) $P(S \leqslant 1.0)$。

5. 如果变量 X 的概率分布如表 5－9 所示。记 $S = X_1 + X_2 + X_3$，其中 X_i 是相互独立且与 X 同分布的随机变量，分别用卷积和矩母函数的方法求 S 的概率分布并确定其期望值。

6. $S = X_1 + X_2 + \cdots + X_6$，$X_i$，$i=1$，2，…，6 是服从伽玛分布的独立同分布随机变量。$E(X_i) = Var(X_i) = i, i=1$，2，…，6。计算 $E[S^3]$。

7. 随机变量 U 的矩母函数为：

$$M_U(t)=(1-2t)^{-9},\ 0<t<\frac{1}{2}$$

表 5-9

x	0	100	200
$f(x)$	0.80	0.15	0.05

(1) 用矩母函数计算 U 的均值和方差。

(2) 用正态函数近似计算 $y_{0.05}$ 和 $y_{0.1}$，其中 y_ε 为满足 $P(U>y_\varepsilon)=\varepsilon$ 的解。

(3) 根据你所掌握的分布函数和矩母函数的信息，确定 U 的分布函数并重新计算 $y_{0.05}$ 和 $y_{0.01}$。

8. 若 $P[\sum_{i=1}^{n}X_i\geqslant n\mu+a\sqrt{n}\sigma]=c+b\Phi(d)$，其中 a 为已知参数，利用中心极限定理求 b，c 和 d。其中 X_i 是独立同分布的随机变量，μ 和 σ 分别为均值和标准差。Φ 是标准正态分布的分布函数。再利用该近似式计算本习题第 4 题中的 (3)。

9. 某火灾保险公司承包了 160 幢建筑物的火灾保险，其保险金额如表 5-10 所示。

表 5-10

保险金额	保险合同数
10 000	80
20 000	35
30 000	25
50 000	15
100 000	5

假设每幢建筑物一年内最多只发生一次火灾，概率为 0.04，再设每幢建筑物的损失变量相互独立，发生损失时的损失额服从 0 到最高保险金额之间的均匀分布。记 N 为一年内保险人的总赔款次数，S 为总赔款额。要求：

(1) 计算 N 的均值和方差；

(2) 计算 S 的均值和方差；

(3) 如果保险人要使承保后总赔款额超过所收保费的概率不超过 1%，计算安全附加费率。(用正态分布近似)

10. 某寿险公司向 2 300 名客户售出一年期短期寿险合同，如表 5-11 所示。该寿险公司作为直接保险人，决定将每份保单超过 1 的损失进行再保险。若再保险人希望收集足够的保费满足实际损失额超过再保费总额的概率不超过 5%，试确定再保险人的安全附加费率。

表 5-11

类别	保险金额（单位化）	死亡概率	保单数目
1	1	0.1	500
2	2	0.02	500
3	3	0.02	300
4	2	0.1	500
5	3	0.1	500

第六章　短期聚合风险模型

学习目标

- □ 了解短期聚合风险模型的内涵以及它与短期个体风险模型的区别和联系
- □ 了解理赔总量模型特别是复合泊松模型的基本性质与特殊性质
- □ 了解聚合理赔量的两个近似模型
- □ 熟悉复合泊松模型的各种性质
- □ 掌握并能运用上述性质解决实际问题

§6.1　引　　言

第五章中讨论的个体风险模型是以每张保单为基本对象，考虑保单组合在一定时期内发生的理赔总量，进而考虑公司的所有保单组合在某段时期的理赔总量。但是，在保单组合给定的保险期间内，大多数保单是不会发生理赔的。因此，可以将保单组合中的保单按照是否发生理赔分为两类，其中不发生理赔的保单不影响保单组合的总损失金额。本章介绍的聚合风险模型是将保单组合视为一个整体，以发生理赔的保单为基本研究对象，理赔总量是按每次理赔发生的时间顺序将所有理赔量累加起来。

用 N 表示某类保单在单位时间（比如一个会计年度）内发生理赔的次数，用 C_i来表示该类保单在此期间第 i 次理赔的金额，则该类保单在此期间的理赔总量 S 可表示为：

$$S=\begin{cases}C_1+C_2+\cdots+C_N=\sum_{i=1}^{N}C_i & N>0\\ 0 & N=0\end{cases}\tag{6.1.1}$$

其中：

1. N 取值为非负整数，而且 $P\{N=0\}>0$，N 是与保单组合的理赔发生频率有关的随机变量，一般称之为**理赔次数变量**；

2. C_i是取值于正数（连续或离散）、测量每次独立理赔量额度大小的随机变量，而且有 $P\{C_i=0\}=0$，一般称之为**理赔额变量**。

为了使模型（6.1.1）在理论上具有可操作性，通常对其有以下的假设：

假设 1　随机变量 N，C_1，C_2，…相互独立。

假设2　C_1，$C_2\cdots$具有相同的分布，即C_i都是同质风险。记它们的共同的概率分布函数为$P(x)$、概率密度（或概率函数）为$p(x)$，用C表示服从该共同分布的随机变量。

独立性假设1是较为普遍的假设，是对实际问题的简化，关于理赔额同分布的假设显然只适用于具有同质风险的同类保单。至此，已完成对模型（6.1.1）的基本说明，在风险理论中一般称模型（6.1.1）为短期聚合风险模型，按照概率论的定义，也称模型（6.1.1）为复合模型或随机和变量。

对于模型（6.1.1）我们主要关心的是聚合理赔量S的分布，也就是研究如何用N的分布和C_i的分布来表示S的分布，所以首先要分析N和C_i的分布。对于N，通常会选择二项分布、泊松分布和负二项分布等离散型分布；对于C_i，通常考虑指数分布、对数正态分布和伽玛分布等取值于正半实轴的连续分布。

如果用泊松分布来描述理赔次数N的分布，则模型S称为**复合泊松分布**，它在风险理论中占有相当重要的地位，我们将在§6.3中详细介绍复合泊松分布及其性质。我们也常用负二项分布来描述理赔次数N的分布，这时称S的分布为**复合负二项分布**，在§6.4中将研究复合泊松分布与复合负二项分布的近似模型。值得一提的是，在本章中矩母函数是一个主要的工具，它对于研究模型（6.1.1）表示的理赔总量提供了很大的帮助。

与个体风险模型不同，组成S的项数N是一个随机变量，虽然理赔额变量相互独立，但是计算聚合理赔额S的期望值和方差必须引入条件期望和条件方差的概念。这些概念不仅在风险理论和精算中，而且在许多应用统计的领域里都起着极端重要的作用。

§6.2　理赔总量模型

在第四章分析理赔次数和理赔额分布的基础上，我们开始讨论理赔总量模型。正如第五章中用个体风险模型描述理赔总量，这里用聚合风险模型描述理赔总量也无非是要获得理赔总量$S=\sum_{i=1}^{N}C_i$的概率分布信息。按顺序可类似地将讨论归结为如下三个问题：

1. 关于S的概率分布和概率密度；
2. 关于S的均值、方差或更高阶矩；
3. 关于S的概率分布的近似和逼近。

在以下两小节中我们首先一般性地讨论问题1和2，本章第三节将利用复合泊松分布讨论问题3。

6.2.1 S 的概率分布

首先从分布函数的定义出发，直接计算聚合模型（6.1.1）的分布函数和分布密度，设 S 的分布函数为 $F_S(x)$，则有

$$\begin{aligned} F_S(x) &= P(S \leqslant x) \\ &= \sum_{n=0}^{\infty} P(S \leqslant x \mid N = n)P(N = n) \qquad (6.2.1) \\ &= P(N = 0) + \sum_{n=1}^{\infty} P(N = n)P\{C_1 + \cdots + C_n \leqslant x\} \end{aligned}$$

由于 C_1，C_2，…，C_n 独立同分布，共同的概率分布函数为 $F(x)$、概率密度函数为 $p(x)$，则

$$P(C_1 + C_2 + \cdots + C_n \leqslant x) = F * F * \cdots * F(x) = F^{*n}(x)$$

将此代入式（6.2.1），并且定义 $F^{*0}(x) \equiv 1$，可得：

$$F_S(x) = \sum_{n=0}^{\infty} P(N = n)F^{*n}(x) \qquad (6.2.2)$$

相应地，也有 S 的密度函数 $f_S(x)$ 的表示：

$$f_S(x) = \sum_{n=0}^{\infty} P(N=n)p^{*n}(x) \qquad (6.2.3)$$

对于离散型随机变量，这里的概率密度理解为 $p(x) = P\{X = x\}$。

【例 6－1】 假设某保单组合在单位时间内至多可发生 3 次理赔，而且已知理赔次数为 0、1、2 和 3 的概率分别为 0.1、0.3、0.4 和 0.2。同时，每次的理赔额为 1、2 或 3 个货币单位，相应的概率分别为 0.5、0.4 和 0.1。试计算理赔总量 S 的概率分布。

解：

已知：$N \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0.1 & 0.3 & 0.4 & 0.2 \end{pmatrix}$，$C \sim \begin{pmatrix} 1 & 2 & 3 \\ 0.5 & 0.4 & 0.1 \end{pmatrix}$

按公式（6.2.3）有

$$\begin{aligned} f_S(x) &= \sum_{n=0}^{3} P\{N=n\}p^{*n}(x) \\ &= 0.1p^{*0}(x) + 0.3p^{*1}(x) + 0.4p^{*2}(x) + 0.2p^{*3}(x) \end{aligned}$$

其中：$p^{*0}(x)$ 为 0 点的退化分布，$p^{*1}(x) = p(x)$，$p^{*2}(x)$ 和 $p^{*3}(x)$ 则需要通过卷积递推计算得到。通过具体的计算，结果列入表 6－1。

表 6－1　　例 6－1 的计算结果

(1)	(2)	(3)	(4)	(5)	(6)	(7)
X	$p^{*0}(x)$	$p^{*1}(x)=p(x)$	$p^{*2}(x)$	$p^{*3}(x)$	$f_S(x)$	$F(x)$
0	1.0	—	—	—	0.1000	0.1000
1	—	0.5	—	—	0.1500	0.2500

续表

(1)	(2)	(3)	(4)	(5)	(6)	(7)
X	$p^{*0}(x)$	$p^{*1}(x)=p(x)$	$p^{*2}(x)$	$p^{*3}(x)$	$f_S(x)$	$F(x)$
2	—	0.4	0.25	—	0.2200	0.4700
3	—	0.1	0.40	0.125	0.2150	0.6850
4	—	—	0.26	0.300	0.1640	0.8490
5	—	—	0.08	0.315	0.0950	0.9440
6	—	—	0.01	0.184	0.0408	0.9848
7	—	—	—	0.063	0.0126	0.9974
8	—	—	—	0.012	0.0024	0.9998
9	—	—	—	0.001	0.0002	1.0000

■

6.2.2　S 的均值、方差或高阶矩

利用命题4－1，将模型（6.1.1）中的 S 和 N 分别代入式（4.3.12）和（4.3.13）中的 X 和 Y，自然得到以下的命题6－1。

命题6－1　若模型（6.1.1）中的 N 和 C 的数学期望和方差都存在，则有

$$E(S)=E(N)E(C) \tag{6.2.4}$$

$$Var(S)=E^2(C)Var(N)+E(N)Var(C) \tag{6.2.5}$$

证明：

（1）直接代入式（4.3.12），有

$$E(S)=E[E(S|N)]=E[NE(C_i)]=E(N)E(C_i)$$

（2）直接代入式（4.3.13），有

$$\begin{aligned}Var(S)&=Var[E(S|N)]+E[Var(S|N)]\\&=Var[NE(C_i)]+E[NVar(C_i)]\\&=E^2(C_i)Var(N)+E(N)Var(C_i)\end{aligned}$$

表达式（6.2.4）表明理赔总量的期望值等于平均理赔次数与平均理赔额的乘积，这与现实的直观感觉是相符合的。表达式（6.2.5）可相应地理解为理赔总量的方差由两部分构成：一部分是理赔量本身的方差，另一部分是理赔次数的方差。

【例6－2】　设理赔次数 N 服从负二项分布，参数 $p=1/3, Var(N)=24$，又知理赔额 $C_i\sim\begin{pmatrix}2&3&4\\0.1&0.4&0.5\end{pmatrix}$，求聚合理赔量 S 的均值与方差

之和。

解： 由已知条件，有 $Var(N)=\frac{rq}{p^2}=6r=24$，可得 $r=4$。因此，

$$E(N)=\frac{rq}{p}=2r=2\times4=8$$

又由
$$E(C_i)=2\times0.1+3\times0.4+4\times0.5=3.4$$
$$Var(C_i)=E(C_i^2)-[E(C_i)]^2$$
$$=4\times0.1+9\times0.4+16\times0.5-3.4^2=0.44$$

因此有

$$E(S)+Var(S)=E(N)E(C_i)+[E(C_i)^2Var(N)+E(N)Var(C_i)]$$
$$=8\times3.4+3.4^2\times24+8\times0.44=308.16$$

下面考虑 S 的矩母函数。按照矩母函数的定义有：

$$M_S(t)=E(e^{tS})=E[E(e^{tS}|N)]$$
$$=E\{[M_{C_i}(t)]^N\}=E[(e^{N\log M_{C_i}(t)})]$$
$$=M_N[\ln M_{C_i}(t)] \tag{6.2.6}$$

因此，只要已知理赔次数 N 的矩母函数 $M_N(t)$ 和理赔额 C_i 的矩母函数 $M_{C_i}(t)$，便可以按照式（6.2.6）将这两个函数进行复合运算得到 S 的矩母函数。■

【例6－3】 假定 N 服从几何分布，理赔额 C_i 服从均值为 1 的指数分布，试求 S 的矩母函数 M_S（t）及分布函数。

解： 由 N 服从几何分布，有

$$P(N=n)=pq^n,\qquad n=0,1,2,\cdots,0<q<1,p=1-q$$

按照矩母函数的定义，有

$$M_N(t)=\frac{p}{1-qe^t},\ 0<t<-\ln(1-p)$$

因此，
$$M_S(t)=\frac{p}{1-qM_C(t)}$$

而
$$M_C(t)=\frac{1}{1-t},\qquad 0<t<1$$

代入式（6.2.6），得：

$$M_S(t)=\frac{p}{1-q\times\frac{1}{1-t}}=p+q\frac{p}{p-t},\qquad 0<t<p \tag{6.2.7}$$

因为 $p/(p-t)$ 是指数分布 $F(x)=1-e^{-px}$，$x>0$ 的矩母函数，所以式（6.2.7）可以看做是对 $x=0$ 点的退化分布和以 p 为参数的指数分布的加权。与此相对应，反解出 S 的分布函数也是上述两个分布的加权平均，权重如式（6.2.7）所示，即

$$F(x)=p\times 1+q\times(1-e^{-px})=1-qe^{-px}=1-(1-p)e^{-px},\quad x\geqslant 0 \tag{6.2.8}$$

式（6.2.8）的分布为混合型分布，$F(0)=p$，分布函数如图 6－1 所示。

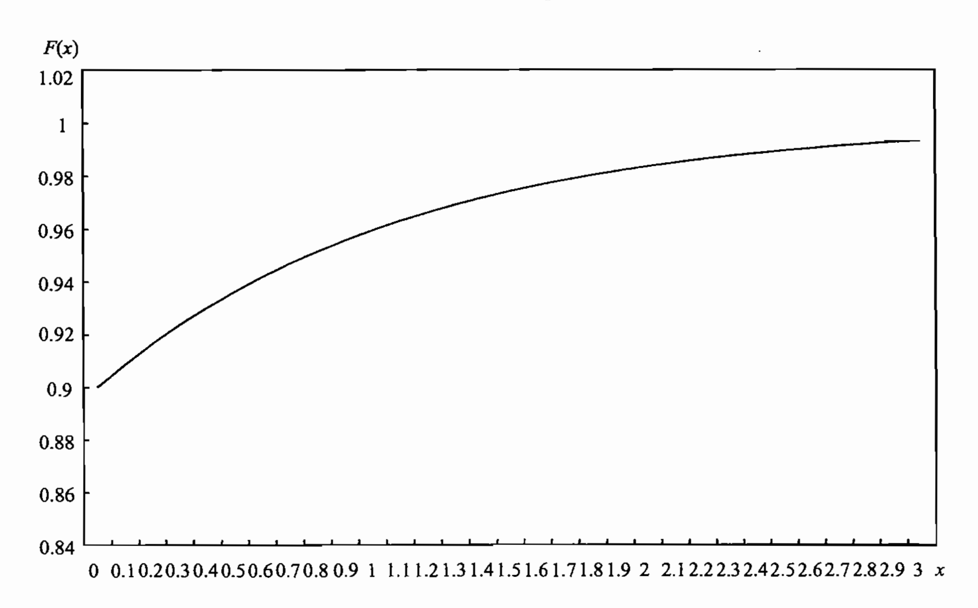


图 6－1　混合指数分布函数（$p=0.9$）

■

6.3　复合泊松模型

如前所述，若 N 服从泊松分布，则称聚合理赔量模型（6.1.1）为**复合泊松模型**。复合泊松模型在风险理论中占有十分重要的地位，被誉为风险理论的经典模型。关于风险模型的许多理论性质都是围绕着复合泊松模型进行的。因此，本节将专门讨论和概括复合泊松模型及其性质。

6.3.1　复合泊松模型的定义和基本性质

为叙述方便，对复合泊松模型重新定义如下：

定义 6－1　称模型（6.1.1）中的随机变量 S 为参数 $\lambda>0$ 的复合泊松模型，若它满足：

（1）N 服从参数为 $\lambda>0$ 的泊松分布；

（2）理赔额变量 C_1，C_2，…互相独立具有相同的分布，简称为理赔额变量 C。其分布函数为 $F(x)$，$x\geqslant 0$；密度函数为 $p(x)$，$x\geqslant 0$；并记其 k 阶原点矩为：

$$p_k=\int_0^\infty x^k dF(x),k=1,2,\cdots$$

(3) N 与 C_1，C_2，…互相独立。

关于复合泊松模型 S 的分布函数和密度，由式（6.2.2）、(6.2.3) 可直接得到：

$$F_S(x) = \sum_{n=0}^{\infty} P(N = n)F^{*n}(x) = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!}F^{*n}(x) \tag{6.3.1}$$

$$f_S(x) = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!}p^{*n}(x) \tag{6.3.2}$$

同时，由式（6.2.4）和（6.2.5）直接有：

$$E(S) = p_1 E(N) = \lambda p_1 \tag{6.3.3}$$

$$Var(S) = \lambda(p_2 - p_1^2) + p_1^2\lambda = \lambda p_2 \tag{6.3.4}$$

又由式（4.3.1）、(4.3.2）并代入式（6.2.6），得到 S 的矩母函数：

$$M_S(t) = M_N[\ln M_C(t)] = e^{\lambda[M_C(t)-1]} \tag{6.3.5}$$

复合泊松模型除了上面的基本性质外，还有很多特殊的性质，这些也是复合泊松模型最有吸引力的性质。

6.3.2 复合泊松模型的特殊性质

我们基本上可以将复合泊松模型的各种特殊性质归纳为：对求和的封闭性、可分解性和分布计算的递推性质。

1. 对求和的封闭性。

定理 6－1 已知 S_1，S_2，…，S_m是相互独立的随机变量。而且 S_i为参数 λ_i的复合泊松分布，理赔额变量的分布函数为 $F_i(x)$，$i=1$，2，…，m，则 $S = S_1 + S_2 + \cdots + S_m$ 服从参数为 $\lambda = \sum_{i=1}^{m} \lambda_i$ 的复合泊松分布，理赔额变量的分布函数为：

$$F(x) = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda}F_i(x) \tag{6.3.6}$$

证明：令 $M_i(t)$表示 $F_i(x)$对应的矩母函数，则 S_i的矩母函数为：

$$M_{S_i}(t) = \exp\{\lambda_i[M_i(t) - 1]\}, \qquad i = 1, 2, \cdots, m$$

由于 S_1，S_2，…，S_m相互独立，S 的矩母函数为：

$$\begin{aligned} M_S(t) &= \prod_{i=1}^{m} M_{S_i}(t) \\ &= \prod_{i=1}^{m} \exp\{\lambda_i[M_i(t) - 1]\} = \exp\left\{\sum_{i=1}^{m} \lambda_i[M_i(t) - 1]\right\} \\ &= \exp\left\{\left(\sum_{i=1}^{m}\lambda_i\right)\left[\sum_{i=1}^{m}\frac{\lambda_i M_i(t)}{\sum_{i=1}^{m}\lambda_i} - 1\right]\right\} = \exp\left\{\lambda\left[\sum_{i=1}^{m}\frac{\lambda_i M_i(t)}{\lambda} - 1\right]\right\} \end{aligned} \tag{6.3.7}$$

其中 $\lambda = \sum_{i=1}^{m} \lambda_i$，而式（6.3.7）可看出 S 的分布是以 $\lambda = \sum_{i=1}^{m} \lambda_i$ 为泊松参

数、理赔额变量分布如式（6.3.6）的复合泊松分布。请读者自行证明分布（6.3.6）的矩母函数为 $\sum_{i=1}^{m}\frac{\lambda_i M_i(t)}{\lambda}$ 所示复合泊松分布的矩母函数。■

该定理对于建立保险风险模型具有两方面的意义：第一，在考虑由多个保单组合构成的总业务组合时，若这些保单组合之间是相互独立的，而且每个保单组合的总理赔模型均为复合泊松模型，则总业务组合的总理赔模型依然为复合泊松模型。第二，在考虑同一保单组合在若干个连续保险年度中的理赔总量的分布时，若每个保险年度的理赔总量都是复合泊松模型而且相互独立，即使它们未必具有相同的分布，这些年的理赔总量也将服从复合泊松模型。

【例 6－4】 已知 $S=S_1+S_2$，且 S_1 与 S_2 为相互独立的两个复合泊松模型，泊松参数分别为 $\lambda_1=2$ 和 $\lambda_2=3$，理赔额变量的分布分别为：

$$C_1\sim\begin{pmatrix}1 & 2\\ 0.6 & 0.4\end{pmatrix},\quad C_2\sim\begin{pmatrix}1 & 2 & 3 & 4\\ 0.1 & 0.3 & 0.5 & 0.1\end{pmatrix}$$

试计算总理赔量 S 的方差以及其理赔额变量 C 的方差。

解：
$$\begin{aligned}Var(S)&=Var(S_1)+Var(S_2)\\&=\lambda_1E(C_1^2)+\lambda_2E(C_2^2)\\&=2\{1\times0.6+2^2\times0.4\}+3\{1\times0.1+2^2\times0.3+3^2\times0.5+4^2\times0.1\}\\&=26.6\end{aligned}$$

由定理 6－1 知 S 服从复合泊松模型，而且：

$$\lambda=\lambda_1+\lambda_2=5$$

设 S 对应的理赔额变量为 $C\sim F(x)$，则

$$p(x)=\frac{\lambda_1}{\lambda}p_1(x)+\frac{\lambda_2}{\lambda}p_2(x)=\frac{2}{5}p_1(x)+\frac{3}{5}p_2(x),\ x=1,\ 2,\ 3,\ 4$$

因此有

$$p(1)=\frac{2}{5}p_1(1)+\frac{3}{5}p_2(1)=0.4\times0.6+0.6\times0.1=0.3$$

$$p(2)=0.4\times0.4+0.6\times0.3=0.34$$

$$p(3)=0.4\times0+0.6\times0.5=0.3$$

$$p(4)=0.4\times0+0.6\times0.1=0.06$$

进而有

$$E(C)=1\times0.3+2\times0.34+3\times0.3+4\times0.06=2.12$$

$$E(C^2)=1\times0.3+2^2\times0.34+3^2\times0.3+4^2\times0.06=5.32$$

$$Var(C)=E(C^2)-(E(C))^2=5.32-2.12^2=0.83$$ ■

2. 可分解性。

定理 6－2 假设随机变量 S 服从复合泊松分布，参数 $\lambda>0$，理赔额为离散随机变量，概率函数为 $\pi_i=P(C=x_i)$，$i=1,\ 2,\ \cdots,\ m$，其中 x_i 表

示理赔额变量的取值。若记 N_i为 S 中取值为 x_i的次数，$i=1$，2，…，m；则有

$$N = N_1 + N_2 + \cdots + N_m \qquad N>0$$

$$S=\begin{cases}0 & N=0\\ x_1N_1+x_2N_2+\cdots+x_mN_m & N>0\end{cases}$$

则以下结论成立：

（1）N_1，N_2，…，N_m互相独立；

（2）N_i服从参数为 $\lambda_i=\lambda\pi_i$的泊松分布，$i=1$，2，…，m。

证明： 随机向量（N_1，N_2，…，N_m）的矩母函数为：

$$M(t_1,\cdots,t_m) = E[\exp(\sum_{i=1}^{m} t_iN_i)], \qquad t_1\geqslant 0, t_2\geqslant 0,\cdots,t_m\geqslant 0$$

由我们在定理 4－3 中介绍过的多项分布的概念，计算可知多项分布的矩母函数为：

$$M(t_1,\cdots,t_m)=(\pi_1e^{t_1}+\pi_2e^{t_2}+\pi_me^{t_m})^n \tag{6.3.8}$$

在本定理中，N 表示试验次数，当 N 取固定的正整数 n 时，（N_1，N_2，…，N_m）服从参数为（n，π_1，…，π_m）多项分布，矩母函数如（6.3.8）所示，对于随机变量 N，可以用全概率公式求出矩母函数：

$$\begin{aligned}E[\exp(\sum_{i=1}^{m}t_iN_i)] &= \sum_{n=0}^{\infty}E[\exp(\sum_{i=1}^{m}t_iN_i)\mid N=n]P(N=n)\\ &= \sum_{n=0}^{\infty}(\pi_1e^{t_1}+\cdots+\pi_me^{t_m})^n\frac{e^{-\lambda}\lambda^n}{n!}\\ &= e^{-\lambda}\sum_{n=0}^{\infty}\frac{[\lambda\sum_{i=1}^{m}\pi_ie^{t_i}]^n}{n!}\\ &= e^{-\lambda}\sum_{n=0}^{\infty}\frac{z^n}{n!}(\text{其中 } z=\lambda\sum_{i=1}^{m}\pi_ie^{t_i}) = e^{-\lambda}e^z\\ &= \exp\{\sum_{i=1}^{m}\lambda\pi_i(e^{t_i}-1)\} = \prod_{i=1}^{m}\exp\{\lambda\pi_i(e^{t_i}-1)\}\end{aligned}$$

上面乘积式中的第 i 项对应着参数为 $\lambda\pi_i$的泊松分布的矩母函数，而且上式对于任意的一组 $t_1\geqslant 0$，$t_2\geqslant 0$，…，$t_m\geqslant 0$ 都成立，因此 N 可以按照结论（1）和（2）的情形进行分解。 ■

下面我们通过两个例子来讨论定理 6－2 的应用。

若保险合同约定理赔按照以下免赔方式进行，保险人的每次理赔额为

$$Y=X-d\mid(X>d)$$

其中参数 d 为免赔额，X 为实际损失额，Y 为保险公司的理赔额变量。如果在某一段固定时间内损失发生次数是参数为 λ 的泊松变量，则由定理 6－2 可知理赔次数是参数为 λp 的泊松变量，其中：

$$p=P\{X>d\}$$

理赔额变量的分布密度是：

$$f_Y(x)=\frac{f_X(x+d)}{p},\qquad x>0$$

因此，理赔总额是泊松参数为 λp 、理赔额变量的分布密度为 $f_Y(x)$ 的复合泊松随机变量。

【例 6－5】 某种一年期的医疗保险损失模型为：理赔次数服从参数为 $\lambda=0.12$ 的泊松分布，实际损失变量相互独立，其分布见表 6－2。保险约定免赔额为 150 元，赔偿限额为 750 元，即赔付时被保险人自付 150 元以内的损失，保险人赔付剩余的损失，但赔付总量不超过 600 元。试分析保险人在保险期间内总赔付的分布。

表 6－2　　实际损失分布　　（单位：50 元）

x	$f_X(x)$	$F_X(x)$	x	$f_X(x)$	$F_X(x)$	x	$f_X(x)$	$F_X(x)$	x	$f_X(x)$	$F_X(x)$
1	0.02	0.02	6	0.04	0.15	11	0.09	0.50	16	0.06	0.89
2	0.02	0.04	7	0.05	0.20	12	0.09	0.59	17	0.04	0.93
3	0.02	0.06	8	0.06	0.26	13	0.09	0.68	18	0.03	0.96
4	0.02	0.08	9	0.07	0.33	14	0.08	0.76	19	0.02	0.98
5	0.03	0.11	10	0.08	0.41	15	0.07	0.83	20	0.02	1.00

解： 由表 6－2 知实际损失额超过 150 元（3 个货币单位）的概率为 $1-F_X(3)=0.94$，按照合同规理赔额变量的分布为：

$$p(x)=\begin{cases}\dfrac{f_X(x+3)}{0.94}, & x=1,2,\cdots,11\\[2ex]\dfrac{1-F_X(14)}{0.94}, & x=12\end{cases}$$

因此，保险人在一年内的总成本服从泊松参数为 $0.12\times0.94=0.1128$ 的复合泊松模型，$p(x)$ 和 $P(x)$ 具体值见表 6－3。

表 6－3　　实际损失分布　　（单位：50 元）

x	$p(x)$	$P(x)$	x	$p(x)$	$P(x)$	x	$p(x)$	$P(x)$
1	0.0212766	0.0212766	5	0.0638298	0.212766	9	0.0957447	0.5638298
2	0.0319149	0.0531915	6	0.0744681	0.2872340	10	0.0957447	0.6595745
3	0.0425532	0.0957447	7	0.0851064	0.3723404	11	0.0851064	0.7446809
4	0.0531915	0.1489362	8	0.0957447	0.4680851	12	0.2553191	1

■

【例 6－6】 已知聚合理赔 S 服从复合泊松模型，泊松参数 $\lambda=0.8$，理

赔额变量的概率函数为：$\begin{pmatrix} 1 & 2 & 3 \\ 0.25 & 0.375 & 0.375 \end{pmatrix}$。试用两种方法分别计算 S 的概率函数 $f_S(x)=P\{S=x\}$ 在 $x=0$，1，…，6 的取值。

解：

方法一：直接用式（6.3.2），有

$$f_S(x) = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} p^{*n}(x) = \sum_{n=0}^{\infty} \frac{0.8^n}{n!} e^{-0.8} p^{*n}(x), \qquad x = 0,1,\cdots,6$$

计算结果如表 6－4 所示。

表 6－4　　方法一的计算结果

x	$p^{*0}(x)$	$p(x)$	$p^{*2}(x)$	$p^{*3}(x)$	$p^{*4}(x)$	$p^{*5}(x)$	$p^{*6}(x)$	$f_S(x)$
0	1							$e^{-0.8}$
1		0.25						$0.2e^{-0.8}$
2		0.375	0.0625					$0.32e^{-0.8}$
3		0.375	0.1875	0.015625				$0.36e^{-0.8}$
4			0.328125	0.070313	0.003906			$0.11e^{-0.8}$
5			0.28125	0.175781	0.023438	0.000977		$0.11e^{-0.8}$
6			0.140625	0.263672	0.076172	0.007324	0.000024	$0.07e^{-0.8}$

方法二：由于理赔额变量的取值仅有 1、2 和 3 三种情况，所以可以将 S 分解为：

$$S=1N_1+2N_2+3N_3$$

由定理 6－2，N_1、N_2和 N_3相互独立且服从参数分别为 $\lambda_1=0.25\times0.8=0.2$，$\lambda_2=0.375\times0.8=0.3$，$\lambda_3=0.375\times0.8=0.3$ 的泊松分布，只需作两次卷积运算即可得到如表 6－5 所示的结果。

表 6－5　　方法二的计算结果

x	$P\{N_1=x\}$	$P\{2N_2=x\}$	$P\{3N_3=x\}$	$P\{N_1+2N_2=x\}$	$f_S(x)=P\{N_1+2N_2+3N_3=x\}$
0	0.818731	0.740818	0.740818	0.606531	0.449329
1	0.163746	—	—	0.121306	0.089866
2	0.016375	0.222245	—	0.194090	0.143785
3	0.001092	—	0.222245	0.037201	0.162358
4	0.000055	0.033337	—	0.030974	0.049906
5	0.000002	—	—	0.005703	0.047960
6	0	0.003334	0.033337	0.003288	0.030923

■

通过例 6－6 可以看出，在计算聚合理赔 S 的分布时，利用复合泊松分布的可分解性可以减少卷积运算的工作量。事实上，利用分布计算的递推性质，还可以大大提高计算的速度并且递推的方法更适用于计算机

编程。

3. 分布计算的递推性质。

定理 6－3　对于复合泊松模型（6.1.1），当理赔额变量 C 取值于正整数，有如下的 $f_S(x)$ 迭代公式：

$$f_S(0)=e^{-\lambda}$$

$$f_S(x)=\frac{\lambda}{x}\sum_{i=1}^{x} ip(i)f_S(x-i),\qquad x=1,2,3,\cdots \tag{6.3.9}$$

证明： 首先考虑如下的条件期望计算：

$$E(C_k|C_1+C_2+\cdots+C_{n+1}=x),\qquad k=1,2,3,\cdots,n+1$$

显然有

$$\sum_{k=1}^{n+1}E(C_k|\sum_{i=1}^{n+1}C_i=x)=x,\qquad x=1,2,3,\cdots \tag{6.3.10}$$

由 C_i同分布的条件，可知式（6.3.10）左边各项都相等，从而有

$$E(C_k|C_1+\cdots+C_{n+1}=x)=\frac{x}{n+1},\qquad x=1,2,3,\cdots$$

按照条件期望的定义，有

$$\begin{aligned}E(C_k\mid\sum_{i=1}^{n+1}C_i=x)&=E(C_{n+1}\mid\sum_{i=1}^{n+1}C_i=x)=\sum_{j=1}^{x}jP\{C_{n+1}=j\mid\sum_{i=1}^{n+1}C_i=x\}\\&=\sum_{j=1}^{x}j\frac{P\{C_{n+1}=j\}P\{\sum_{i=1}^{n}C_i=x-j\}}{P\{\sum_{i=1}^{n+1}C_i=x\}}\\&=\sum_{j=1}^{x}jp(j)\frac{p^{*n}(x-j)}{p^{*n+1}(x)},\qquad x=1,2,3,\cdots\end{aligned}$$

因此有

$$\frac{x}{n+1}p^{*n+1}(x)=\sum_{i=1}^{x}ip(i)p^{*n}(x-i),\qquad x=1,2,3,\cdots$$

因此，当 $x=1,2,3,\cdots$ 时，按照定义直接有

$$\begin{aligned}f_S(x)&=P\{\sum_{k=1}^{N}C_k=x\}\\&=\sum_{n=0}^{\infty}P\{N=n+1\}P\{\sum_{k=1}^{n+1}C_k=x\}=\sum_{n=0}^{\infty}\frac{\lambda^{n+1}e^{-\lambda}}{(n+1)!}\times p^{*n+1}(x)\\&=\sum_{n=0}^{\infty}\frac{\lambda^{n+1}e^{-\lambda}}{(n+1)!}\times\frac{n+1}{x}\sum_{i=1}^{x}ip(i)p^{*n}(x-i)=\sum_{i=1}^{x}\frac{i\lambda p(i)}{x}\sum_{n=0}^{\infty}\frac{\lambda^{n}e^{-\lambda}}{n!}p^{*n}(x-i)\\&=\frac{\lambda}{x}\sum_{i=1}^{x}ip(i)\,f_S(x-i)\end{aligned}$$

■

【例 6－7】　试用定理 6－3 的迭代方法重新计算例 6－6。

解： 由已知 $\lambda=0.8$，$f_S(0)=e^{-0.8}$，由式（6.3.9），有

$$f_S(1)=\frac{0.8}{1}p(1)f_S(0)=0.8\times0.25\times f_S(0)=0.2e^{-0.8}$$

$$f_S(2)=\frac{0.8}{2}\{0.25f_S(1)+2\times0.375f_S(0)\}=0.32e^{-0.8}$$

$$f_S(3)=\frac{0.8}{3}\{0.25f_S(2)+2\times0.375f_S(1)+3\times0.375f_S(0)\}$$
$$=0.36e^{-0.8}$$

……

其结果与例 6－6 相同的结果，具体见表 6－5。

【例 6－8】 设某保险人的总风险服从复合泊松模型，年平均理赔次数为 0.2 次。在任何一次理赔中，有 80% 的概率会损失 1 万元，20% 的概率会损失 2 万元。试计算保险人总损失的概率分布。

解： 为简化表达，我们用 1 万元表示单位货币，因此有

$$\lambda=0.2, p(1)=0.8=1-p(2)$$

根据递推公式（6.3.9），可得：

$$f_S(0)=e^{-\lambda}=e^{-0.2}=0.818731$$

$$f_S(1)=\lambda p(1)f_S(0)=0.2\times0.8\times0.818731=0.130997$$

$$f_S(2)=\frac{\lambda}{2}\{p(1)f_S(1)+2p(2)f_S(0)\}=0.043229$$

$$f_S(3)=\frac{\lambda}{3}\{p(1)f_S(2)+2p(2)f_S(1)\}=0.005799$$

如此迭代下去，结果列入表 6－6。

表 6－6　总损失的概率分布

x	$f_S(x)$	$F_S(x)$
0	0.818731	0.818731
1	0.130997	0.949728
2	0.043229	0.992957
3	0.005799	0.998755
4	0.001097	0.999852
5	0.000128	0.999980
6	0.000018	0.999998

■

4. 递推性质的应用——限额损失再保险问题。最常见的两种最基本的再保险方式为：限额损失再保险（也称超额超赔再保险）和比例再保险。若将直接承保的聚合理赔记为 S，则上述两种再保险模型可分别表示如下：

限额损失再保险：

$$I_d(S)=\begin{cases}0, & S\leqslant d\\ S-d, & S>d\end{cases}\tag{6.3.11}$$

损失比例再保险：$I(S)=kS,\quad 0<k<1$　　(6.3.12)

有时，在不会产生误解的前提下将 $I_d(S)$ 简记为 I_d。正如第四章中所强调的，限额损失再保险模型无论在理论上还是在实际中都具有特别的重要性，因此是讨论的重点。式（6.3.11）中的 d 同样称为免赔额，对原保险人来说，d 起到了“限制损失”的作用，因为超过额度 d 的部分即 I_d 就是再保险人承担的风险，而原保险人自留的风险则是：

$$S - I_d(S) = \begin{cases} S, & S \leqslant d \\ d, & S > d \end{cases} \qquad (6.3.13)$$

即 S 被分解为 $S = (S - I_d) + I_d$。

显然，无论对原保险人还是对再保险人来说都必须研究最佳的 d 是什么，即自留多少、分保多少的问题。更为基础的问题是：限额损失再保险理赔额 I_d的均值，即纯保费 $E(I_d)$。设 S 的分布函数为 $F_S(x)$，密度函数为 $f_S(x)$，则有

$$\begin{aligned} E(I_d) &= \int_d^{\infty} (x - d) f_S(x) dx \\ &= E(S) - d + \int_0^d (d - x) f(x) dx \qquad (6.3.14) \\ &= \int_d^{\infty} [1 - F_S(x)] dx \\ &= E(S) - \int_0^d [1 - F_S(x)] dx \qquad (6.3.15) \end{aligned}$$

特别地，当理赔 S 仅取非负整数值并且 d 也是整数时，有

$$\begin{aligned} E(I_d) &= \sum_{x=d+1}^{\infty} (x - d) f_S(x) \\ &= f_S(d+1) + 2f_S(d+2) + \cdots \qquad (6.3.16) \\ &= E(S) - d + \sum_{x=0}^{d-1} (d - x) f_S(x) \qquad (6.3.17) \\ &= \sum_{x=d}^{\infty} [1 - F_S(x)] \\ &= [1 - F_S(d)] + [1 - F_S(d+1)] + \cdots \qquad (6.3.18) \\ &= E(S) - \sum_{x=0}^{d-1} [1 - F_S(x)] \qquad (6.3.19) \end{aligned}$$

由这些不同的表达式还可以导出关于 $E(I_d)$ 的递推公式：

$$E(I_{d+1}) = E(I_d) - [1 - F_S(d)],\ d = 0, 1, 2, \cdots,\ E(I_0) = E(S) \qquad (6.3.20)$$

【例 6-9】 设 $S = \begin{cases} \sum_{i=1}^{N} X_i, & N > 0 \\ 0, & N = 0 \end{cases}$，其中理赔次数 $N \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0.1 & 0.3 & 0.4 & 0.2 \end{pmatrix}$，理赔额变量 $X \sim \begin{pmatrix} 1 & 2 & 3 \\ 0.5 & 0.4 & 0.1 \end{pmatrix}$，试计算 $E(I_7)$。

解： 因为 N 的最大值为 3，S 的最大值为 9，所以用公式（6.3.16）计算 $E(I_7)$ 比较简洁。

$$\begin{aligned} f_S(8) &= P\left(\sum_{i=1}^{3} X_i = 8 \mid N = 3\right) P(N = 3) \\ &= \binom{3}{2} P(X_1 = 3) P(X_2 = 3) P(X_3 = 2) P(N = 3) \end{aligned}$$

$$= \frac{3\times 2}{2\times 1}0.1^2\times 0.4\times 0.2 = 0.0024$$

$$f_S(9) = P(\sum_{i=1}^{3} X_i = 9 \mid N = 3)P(N = 3)$$
$$= P(X_1 = 3)P(X_2 = 3)P(X_3 = 3)P(N = 3)$$
$$= 0.1^3\times 0.2 = 0.0002$$

因此有

$$E(I_7) = f_S(8) + 2f_S(9) = 2.8\times 10^{-3}$$ ■

【例 6－10】 设聚合理赔 S 为复合泊松分布，参数 $\lambda = 1.5$，理赔额变量的分布是 $p(1) = \frac{2}{3} = 1 - p(2)$，试计算 $f_S(x)$ 、$F_S(x)$和 $E(I_x)$,$x = 1, 2, \cdots, 10$。

解：因为理赔额变量 X 的取值只有整数 1 和 2，所以聚合理赔 S 的取值也是非负整数，而且，一、二阶原点矩为：

$$p_1 = 1\times\frac{2}{3} + 2\times\frac{1}{3} = \frac{4}{3}, p_2 = 1\times\frac{2}{3} + 2^2\times\frac{1}{3} = 2$$

由复合泊松分布的迭代公式$f_S(x) = \frac{\lambda}{x}\sum_{y=1}^{x} yp(y)f_S(x-y)$，有

$$f_S(0) = e^{-1.5}, f_S(1) = 1.5\times\frac{2}{3}\times f_S(0) = e^{-1.5}$$

$$f_S(x) = \frac{1.5}{x}\left[\frac{2}{3}f_S(x-1) + \frac{2}{3}f_S(x-2)\right] = \frac{1}{x}[f_S(x-1) + f_S(x-2)], \quad x = 2,3,\cdots$$

$$F_S(x) = F_S(x-1) + f_S(x), \quad x = 1,2,\cdots$$

因此有

$$E(I_x) = E(I_{x-1}) - [1 - F_S(x-1)], \quad x = 1, 2, \cdots$$

$$E(I_0) = E(S) = \lambda p_1 = 2$$

具体计算结果（保留到小数点后 3 位）见表 6－7。

表 6－7　　例 6－10 的计算结果

x	$f_S(x)$	$F_S(x)$	$E(I_x)$
0	0.223	0.223	2.00
1	0.223	0.446	1.223
2	0.223	0.669	0.669
3	0.149	0.818	0.339
4	0.093	0.911	0.157
5	0.048	0.959	0.068
6	0.024	0.983	0.027
7	0.010	0.993	0.011
8	0.004	0.997	0.004
9	0.002	0.999	0.001
10	0.001	1.000	0.000

■

从例 6－10 可以发现，随着自留额的上升，再保险的风险逐渐降低，而且，有时可能对于不是很大的自留额（如本例中的 $x=10$）就会使再保险的平均损失接近于零。

5. 递推性质的拓展。定理 6－3 的迭代公式不仅对复合泊松分布成立，对于如二项分布、负二项分布、几何分布等其他分布也可导出类似的公式，实际上，类似式（6.3.9）的递推公式对（a，b，0）类计数随机变量都成立。

定理 6－4　若模型（6.1.1）中的理赔次数 N 服从（a，b，0）类计数分布，而且理赔额变量 C 取值于正整数，则有如下关于理赔总额概率分布函数 $f_S(x)$ 的迭代公式：

$$f_S(0)=P(N=0)$$

$$f_S(x)=\sum_{i=1}^{x}\left(a+\frac{bi}{x}\right)p(i)f_S(x-i),\qquad x=1,2,3,\cdots \tag{6.3.21}$$

定理 6－4 的证明与定理 6－3 的证明类似，这里不再重复。

§6.4　聚合理赔量的近似模型

在前三节我们讨论了总理赔模型（6.1.1）的精确分布，我们现在就复合泊松分布和复合负二项分布的情况讨论聚合理赔的近似分布。下面分两种情形讨论：在理赔分布基本上对称的情形采用正态近似，在理赔分布有偏斜的情形采用平移伽玛分布近似。

6.4.1　正态分布近似

定理 6－5

（1）当模型（6.1.1）为复合泊松模型、泊松参数为 λ、理赔额变量的分布函数为 $F(x)$ 时，有

$$Z=\frac{S-\lambda p_1}{\sqrt{\lambda p_2}}$$

的分布当 $\lambda\to\infty$ 时趋于标准正态分布，其中 p_1、p_2 分别为 $F(x)$ 对应的 1 阶和 2 阶原点矩。

（2）当模型（6.1.1）为复合负二项分布，参数为 r 和 p，理赔额变量的分布函数为 $F(x)$ 时，

$$Z=\frac{S-r\left(\frac{q}{p}\right)p_1}{\sqrt{r\left(\frac{q}{p}\right)p_2+r\left(\frac{q}{p}\right)^2p_1^2}}$$

的分布当 $r\to\infty$ 时趋于标准正态分布。

证明：

（1）对于泊松分布的情形，我们将通过证明 $M_Z(t)$ 的极限为 $e^{\frac{t^2}{2}}$ 来证明定理的结论。

由 $Z=\dfrac{S-\lambda p_1}{\sqrt{\lambda p_2}}$ 可得：

$$M_Z(t)=M_S\left(\frac{t}{\sqrt{\lambda p_2}}\right)\exp\left[-\frac{\lambda p_1 t}{\sqrt{\lambda p_2}}\right]$$

又由式（6.3.5）得：

$$M_Z(t)=\exp\left\{\lambda\left[M_C\left(\frac{t}{\sqrt{\lambda p_2}}\right)-1\right]-\frac{\lambda p_1 t}{\sqrt{\lambda p_2}}\right\}$$

再由 $M_C(t)=1+\dfrac{p_1}{1!}t+\dfrac{p_2}{2!}t^2+\cdots$，得

$$M_Z(t)=\exp\left[\frac{1}{2}t^2+o\left(\frac{1}{\sqrt{\lambda}}\right)\right],t>0$$

当 $\lambda\to\infty$ 时，$M_Z(t)\to e^{\frac{t^2}{2}}$，即 $\lim\limits_{\lambda\to\infty}M_Z(t)=e^{\frac{t^2}{2}}$。

（2）对于负二项分布的情形，证明过程类似，不再赘述。 ■

6.4.2 平移伽玛分布近似

- 若 S 为复合泊松模型，其三阶矩为：

$$E[(S-E(S))^3]=\lambda p_3$$

显然不一定为0；

- 若 S 为复合负二项分布，其三阶矩为：

$$E[(S-E(S))^3]=\frac{rqp_3}{p}+\frac{3rq^2p_1p_2}{p^2}+\frac{2rq^3p_1^3}{p^3}$$

显然也不一定为0。因此，这时 S 的分布是有偏斜的。由于正态随机变量的分布是对称的，因此当 S 的偏度较大时使用正态近似并不合适，如果考虑选用平移伽玛分布进行近似，可能更为恰当。

这里我们用 $Gamma\ (x;\ \alpha,\ \beta)$ 表示参数为 α 和 β 的伽玛随机变量的分布函数，即

$$Gamma\ (x;\ \alpha,\ \beta)=\int_0^x\frac{t^{\alpha-1}e^{-\frac{t}{\beta}}}{\Gamma(\alpha)\beta^\alpha}dt,\qquad x\geqslant 0$$

对任意的点 $x_0>0$，定义如下新的分布函数，称其对应的随机变量服从平移伽玛分布：

$$H(x;\alpha,\beta,x_0)=Gamma(x-x_0;\alpha,\beta),\qquad x\geqslant x_0 \tag{6.4.1}$$

这个分布相当于将分布 $Gamma(x;\alpha,\beta)$ 平移了 x_0 个单位。

平移伽玛分布包含参数 α、β 和 x_0，通过简单的推导易知平移伽玛分

布（这里也用 H 代表该分布对应的随机变量）的一、二及三阶中心矩分别为：

$$E(H)=x_0+\alpha\beta, Var(H)=\alpha\beta^2, E\{[H-E(H)]^3\}$$
$$=2\alpha\beta^3 \tag{6.4.2}$$

在用平移伽玛分布来近似聚合理赔 S 时，我们采用中心矩相等的原则，令平移伽玛随机变量和聚合理赔 S 的一、二、三阶中心矩相等，因此有如下三个关于近似平移伽玛分布的参数方程：

$$E(S)=x_0+\alpha\beta, Var(S)=\alpha\beta^2, E\{[S-E(S)]^3\}=2\alpha\beta^3 \tag{6.4.3}$$

解得：

$$x_0=E(S)-2\frac{[Var(S)]^2}{E\{[S-E(S)]^3\}}$$
$$\alpha=4\frac{[Var(S)]^3}{\{E[S-E(S)]^3\}^2},\qquad \beta=\frac{1}{2}\frac{E\{[S-E(S)]^3\}}{Var(S)} \tag{6.4.4}$$

特别地，对于复合泊松分布，有

$$x_0=\lambda p_1-2\lambda\frac{p_2^2}{p_3},\qquad \alpha=4\lambda\frac{p_2^3}{p_3^2},\qquad \beta=\frac{p_3}{2p_2} \tag{6.4.5}$$

再回到式（6.4.3），若取参数 β 的一个从右面趋于 0 的序列 $\beta_n\to 0$，然后令：

$$x_0=\mu-\frac{\sigma^2}{\beta_n},\qquad \alpha=\frac{\sigma^2}{\beta_n^2},\qquad \text{其中 }\mu\in(-\infty,+\infty),\ \sigma>0\text{ 为常数}$$

则当 $\beta_n\to 0$ 时，有：$x_0\to-\infty$，$\alpha\to\infty$，分布 $H(x:\alpha,\beta,x_0)$ 趋于 $N(\mu,\sigma^2)$。因此，正态分布可以看做是这种三个参数分布的一种极限情况，从这个意义上说，平移伽玛分布近似是正态近似的推广情况。

【例 6－11】 考虑某参数 $\lambda=16$ 的泊松分布，试用平移伽玛分布和正态分布近似该分布。

解： 泊松分布随机变量 N 可看做是理赔额变量 C 退化为仅在 1 个货币单位点取值的一种复合泊松分布，即 $N=\sum_{i=1}^{N}C_i, P(C_i=1)=1, i=1,2,\cdots$，因此 C 的各阶矩为：

$$p_k=1,\qquad k=1,\ 2,\ 3$$

由式（6.4.5），则有

$$x_0=-\lambda,\qquad \alpha=4\lambda,\qquad \beta=2$$

解得：$x_0=-16$，$\alpha=64$，$\beta=2$。

若采用正态近似，参数为：

$$\mu=\lambda=16,\ \sigma=\sqrt{\lambda}=4$$

因为 N 的分布函数有解析表达式，可以直接计算，所以表 6－8 比较了三个分布在 $x=5$，10，15，…，40 的分布函数结果。需要注意的是，表 6－8

对两个连续近似分布在各点的值按照中点值进行了修正。

从表 6－8 可以看出，平移伽玛分布的近似效果显然要比正态分布好得多。

表 6－8　　用平移伽玛分布和正态分布近似泊松分布

精确分布（泊松分布）		近似分布	
x	$e^{-16}\sum_{k=0}^{x}\frac{16^k}{k!}$	$Gamma(x+16.5;64,2)$	$\Phi\left(\frac{x+0.5-16}{4}\right)$
5	0.001384	0.001636	0.004332
10	0.077396	0.077739	0.084566
15	0.466745	0.466560	0.450262
20	0.868168	0.868093	0.869705
25	0.986881	0.986604	0.991226
30	0.999433	0.999378	0.999856
35	0.999988	0.999985	0.999999
40	1.000000	1.0000000	1.000000

■

【例 6－12】 假设 S 服从复合泊松分布，参数 $\lambda=12$，且理赔额变量 X 在 [0，1] 上均匀分布，试分别用正态近似和平移伽玛近似计算 $P(S<10)$。

解： 经过简单的计算，易知：$p_1=\frac{1}{2}$，$p_2=\frac{1}{3}$，$p_3=\frac{1}{4}$；$E(S)=\lambda p_1=6$，$Var(S)=\lambda p_2=4$，$E[S-E(S)]^3=\lambda p_3=3$

若采用正态近似，有

$$P(S<10)=P(\frac{S-6}{2}<2)=\Phi(2)=0.9772$$

若采用平移伽玛近似，由式（6.4.3）有

$$\begin{cases}6=x_0+\alpha\beta\\4=\alpha\beta^2\\3=2\alpha\beta^3\end{cases}$$

解得：

$$\begin{cases}\alpha=256/9\\\beta=3/8\\x_0=-14/3\end{cases}$$

因此，　$P(S<10)=Gamma\left(\frac{44}{3};\frac{256}{9},\frac{3}{8}\right)=0.9682$　■

§6.5　个体风险模型与复合泊松模型的关系

第五章中的个体风险模型是以每张保单为基本元素，而本章的聚合风险模型则是以每次理赔为基本对象，并且一般考虑复合泊松模型。虽然考虑的模型不同，但是问题的背景是相同的，所以可以考虑用复合泊松模型近似个体风险模型，并分析两个模型之间的关系。

现有由 n 张独立的保单组成的保单组合，每张保单至多发生一次理赔，而且第 i 张保单发生理赔的概率为：$P(I_i = 1) = q_i$；第 i 张保单发生理赔后的金额为：$B_i \sim f_i(x), x > 0$，相应地，总理赔 S 的模型为：

$$S = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} I_i B_i \tag{6.5.1}$$

若记　$\mu_i = E(B_i), \sigma_i^2 = Var(B_i)$

则有　$$E(S) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} E(I_i B_i) = \sum_{i=1}^{n} \mu_i q_i \tag{6.5.2}$$

$$Var[S] = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} [q_i \sigma_i^2 + \mu_i^2 q_i (1 - q_i)] \tag{6.5.3}$$

下面考虑从两个方面用复合泊松模型近似式（6.5.1）的个体风险模型。

1. 我们以发生总理赔额的均值不变为原则进行复合泊松近似。由于

$$E\left(\sum_{i=1}^{n} I_i\right) = \sum_{i=1}^{n} q_i$$

若以 $\lambda = \sum_{i=1}^{n} q_i$ 作为泊松分布的泊松参数，考虑服从以下分布的理赔额变量 B：

$$B \sim f(x) = \frac{1}{\lambda} \sum_{i=1}^{n} q_i f_i(x) \tag{6.5.4}$$

考虑如下的复合泊松模型：

$$S = \begin{cases} \sum_{i=1}^{N} Y_i & N > 0 \\ 0 & N = 0 \end{cases} \tag{6.5.5}$$

其中：N 为参数 $\lambda = \sum_{i=1}^{n} q_i$ 的泊松分布，Y_i 独立同分布，共同分布如式（6.5.4）所示。则式（6.5.5）可看做是（6.5.1）的一种泊松模型近似，而且有

$$E(S) = \lambda E(B) = \sum_{i=1}^{n} \mu_i q_i,\ Var(S) = \lambda E(B^2) = \sum_{i=1}^{n} q_i (\mu_i^2 + \sigma_i^2) \tag{6.5.6}$$

则式（6.5.5）与（6.5.1）的均值相同，但方差偏高，而且两个方差的差距为：

$$\sum_{i=1}^{n} q_i(\mu_i^2 + \sigma_i^2) - \sum_{i=1}^{n}[q_i\sigma_i^2 + \mu_i^2 q_i(1-q_i)] = \sum_{i=1}^{n}\mu_i^2 q_i^2 > 0,$$

对于理赔额变量 B，其一阶、二阶原点矩还可表达为：

$$E(B) = \sum_{i=1}^{n}\frac{q_i}{\lambda}\times\mu_i = \sum_{i=1}^{n}\lambda_i\mu_i$$

$$E(B^2) = \sum_{i=1}^{n}\frac{q_i}{\lambda}\times(\mu_i^2+\sigma_i^2) = \sum_{i=1}^{n}\lambda_i(\mu_i^2+\sigma_i^2)$$

其中，$\lambda_i = \frac{q_i}{\lambda}$，满足 $\sum_{i=1}^{n}\lambda_i = 1$，对照式（6.5.6），可知在以式（6.5.5）近似的复合泊松模型中，理赔额的均值和二阶原点矩可看做是对每一张保单理赔额的一阶、二阶原点矩的加权平均。

2. 我们以不发生理赔的概率不变为原则进行复合泊松近似。对于参数为 $\tilde{\lambda}$ 的泊松分布 N，有

$$P(N=0) = e^{-\tilde{\lambda}}$$

对于模型（6.5.1），

$$P(\sum_{i=1}^{n} I_i = 0) = (1-q_1)\cdots(1-q_n)$$

若令 $$e^{-\tilde{\lambda}} = (1-q_1)\cdots(1-q_n) \tag{6.5.7}$$

则有 $$\tilde{\lambda} = -\sum_{i=1}^{n}\ln(1-q_i) = \sum_{i=1}^{n}\tilde{\lambda}_i \tag{6.5.8}$$

其中，$\tilde{\lambda}_i = -\ln(1-q_i) = q_i + \frac{q_i^2}{2} + \cdots > q_i$

这时的近似模型仍然为式（6.5.5），理赔额分布仍然为式（6.5.4），只是泊松分布 N 的参数为 $\tilde{\lambda}$。因此，在由式（6.5.8）得到的近似模型（6.5.5）中，S 的均值和方差都比式（6.5.6）偏高：

$$E(S) = \tilde{\lambda}E(B) = \sum_{i=1}^{n}\tilde{\lambda}_i\times\mu_i > \sum_{i=1}^{n}q_i\mu_i$$

$$\begin{aligned} Var(S) &= \tilde{\lambda}E(B^2) \\ &= \sum_{i=1}^{n}\tilde{\lambda}_i(\mu_i^2+\sigma_i^2) > \sum_{i=1}^{n}q_i(\mu_i^2+\sigma_i^2) > \sum_{i=1}^{n}[q_i\sigma_i^2+\mu_i^2 q_i(1-q_i)] \end{aligned}$$

【例 6－13】 现有由 $n=100$ 张独立的保单组成的保单组合，每张保单至多发生一次理赔，而且发生理赔的概率为 $q=0.05$，理赔金额服从相同的分布，记做 $B\sim\begin{pmatrix}1 & 2\\ 0.8 & 0.2\end{pmatrix}$。试讨论对理赔总量 S 的分布近似。

解：按照个体风险模型，理赔总量为：

$$S = \sum_{i=1}^{100} X_i = \sum_{i=1}^{100} I_i B$$

$$P\{I_i = 1\} = 0.05 = 1 - P\{I_i = 0\}$$

由于　$E(X_i) = 0.05 \times E(B) = 0.05 \times 1.2 = 0.06$

$$Var(X_i) = q(1-q)E^2(B) + qVar(B) = 0.0764$$

因此有　$E(S) = 100E(X) = 6, Var(S) = 100Var(X) = 7.64$

按照聚合风险模型的观点，设 N 为理赔次数，则 $N = \sum_{i=1}^{100} I_i$，理赔额的分布还是 B_i的分布。这里的 N 显然不是泊松分布而是一个二项分布：

$$E(N) = nq = 5, Var(N) = npq = 4.75$$

设 S 的概率函数为 $f_S(x) = P\{S = x\}$，由式（6.2.3）知：

$$f_S(x) = \sum_{n=0}^{\infty} P\{N = n\} p^{*n}(x)$$

其中 $p^{*n}(x)$ 是 B 的概率函数的 n 次卷积，要算出它的最后表达式是比较复杂的，在第二章中我们讨论过它的正态近似，但由于这个函数是非负的，通常向右偏斜，正态近似往往低估了它落入均值右边尾部的概率。我们现在考虑用复合泊松模型来近似 $f_S(x)$。

首先按照式（6.5.6）进行近似，泊松参数为 $\lambda = nq = 5$，$p(x)$ 为 B_i的概率函数，按照复合泊松模型的性质，S 的泊松近似 $\tilde{S}$ 有：

$$E(\tilde{S}) = \lambda E(B_i) = 6, Var(\tilde{S}) = \lambda E(B_i^2) = 8$$

均值相同，但复合泊松模型对方差的估计较大，显得比较“保守”。

然后按照式（6.5.8）进行近似，$\tilde{\lambda}_i = -\ln(1-q_i) = -\ln 0.95 = 0.05129, \tilde{\lambda} = 5.129$，

$$E[\tilde{S}] = \tilde{\lambda} E(B_i) = 5.129 \times 1.2 = 6.155$$

$$Var(\tilde{S}) = \tilde{\lambda} E(B_i^2) = 5.129 \times 1.6 = 8.207$$

均高于前一种近似方法的结果。

习　题

1. 若 S 是某地区 1 年内的降雨量之和，试用随机和模型刻画 S。

2. 如果短期聚合风险模型中的理赔次数 N 服从二项式分布 $B(n,p)$，其中的参数 p 服从[0，1]的均匀分布，利用全概率公式计算：（1）N 的矩母函数；（2）N 的均值；（3）N 的方差。

3. 证明：如果 N 服从参数为 λ 的泊松分布，那么，当 $\lambda \to \infty$ 时，变

量 $Z=(N-\lambda)/\sqrt{\lambda}$ 的分布趋向标准正态分布。

4. 如果例 6－1 中的理赔次数 N 服从参数为 6 的泊松分布，其他条件不变，求聚合理赔量 S 的均值和方差。

5. 若聚合理赔模型中理赔额变量服从正态分布 $N(100,9)$，理赔次数 N 的分布如下所示：

$$N\sim\begin{pmatrix}0 & 1 & 2 & 3\\ 0.5 & 0.2 & 0.2 & 0.1\end{pmatrix}$$

求聚合理赔款 S 的均值和方差。

6. 已知聚合理赔模型中 N 服从负二项分布，参数 $r=2$，$p=0.2$。又已知 S 的均值和方差分别为 12.80 和 105.92。试确定理赔额变量的均值和方差。

7. 如果第 6 题中的理赔额变量的分布函数如表 6－9 所示，计算 $f_S(3)$。

表 6－9

x	1	2	3
$P(X=x)$	0.50	0.40	0.10

8. 已知聚合理赔量 S 的理赔次数变量服从概率分布 $P(N=n)=\binom{n+2}{n}\left(\frac{1}{2}\right)^{n+3}$，$n=0,1,2\cdots$，赔款额变量服从均值为 1 的指数分布。(1) 求 S 的矩母函数。(2) 利用矩母函数性质求 S 的均值和方差。

9. 如果 S_1 服从复合泊松模型：参数 $\lambda_1=3, p(1)=p(2)=p(3)=1/3$，$S_2$ 服从参数 $\lambda_2=2$，$p(1)=p(2)=1/2$ 的复合泊松模型，S_1 和 S_2 相互独立。试计算 S_1+S_2 中理赔额变量的分布。

10. 如果聚合理赔量 S 为复合泊松分布模型，参数 $\lambda=0.06$，理赔额变量取 1、2 和 3 的概率分别为 0.20、0.30 和 0.50。试计算 S 不小于 3 的概率。

11. 如果复合泊松模型 S 可以表示为：$S=1N_1+2N_2+3N_3$，并且：

$$E(S)=56,\ Var(S)=126,\ E[(S-E[S])^3]=314$$

求：$\lambda_1=E[N_1]$，$\lambda_2=E[N_2]$，$\lambda_3=E[N_3]$。

12. 如果 S 服从复合泊松模型，参数 $\lambda=5$，已知理赔额变量只取 1 或 2，且 S 的分布函数如表 6－10 所示。利用例 6－6 的两种方法和定理 6－3 计算理赔额变量的分布。

13. 仿照定理 6－3 的证明过程证明定理 6－4。

14. 如果例 6－6 中的聚合理赔 S 服从复合负二项分布，参数 $r=8$，$p=0.4$，其他条件不变。利用定理 6－4 重新计算 S 的概率分布函数。

15. 如果某保险标的一年内的损失次数服从参数为 λ 的泊松分布，损

失额服从指数分布 $f(x)=\beta e^{-\beta x}$，$x>0$。当损失发生时保险人将赔付损失超过 d 的部分。试计算：

（1）一年内总赔款额 S 的分布。

（2）S 的均值和方差。

表 6－10

x	0	1	2	…
$f_S(x)$	e^{-5}	$3.5e^{-5}$	$7.625e^{-5}$	…

16. 已知两类汽车一年内的损失额分别服从参数 $\lambda_1=0.3$ 和 $\lambda_2=0.2$ 的复合泊松分布，理赔额变量的分布如表 6－11 所示。按照条款，两类汽车的免赔额分别为 $d_1=100$，$d_2=150$。求两类汽车一年内总赔款额变量的矩母函数、均值和方差。

表 6－11

x	$f_1(x)$	$f_2(x)$
0	0.20	0.05
50	0.30	0.20
200	0.30	0.50
500	0.10	0.15
1 000	0.10	0.10

17. 对于聚合理赔模型（6.1.1）有：（1）N 服从均值为 0.5 的泊松分布；（2）理赔额变量的均值和方差均为 100。定义损失率为一保单组合的总赔款和总保费的比率，已知相对安全附加费率为 0.1，试用正态近似的方法计算该模型的损失率超过 0.75 的概率。

18. 用平移伽玛分布近似方法估计聚合理赔款 S 的分布，已知：$x_0=0$，$E[S]=\mu$，$Var(S)=\sigma^2$，$E[(S-\mu)^3]=v^3$。试用 μ 和 σ 表示 v。

19. 假设 S 服从参数 $r=6$ 及 $p=1/4$ 的复合负二项分布，理赔额服从指数分布（密度函数为：$3e^{-3x}$，$x>0$），试分别用正态近似和平移伽玛分布计算 $P\{S<8\}$。

20. 分别按照个别理赔模型和两种近似的复合泊松模型对例 6－13 计算 $F_S(x)=P\{S\leqslant 4\}$。

21. 某保险公司对 2 250 个投保人提供医疗费用保险，其分布如表 6－12 所示。每张保单最多理赔一次，总理赔额可以用复合泊松分布近似。拟合的复合泊松分布和个别风险模型的期望损失次数相同，发生损失时个别理赔额的分布也相同。试用复合泊松分布模型求理赔总量方差的近似值。

表 6－12

保单类别	被保险人数	理赔概率	理赔分布
1	1 500	0.01	在（0，1）上均匀分布
2	750	0.02	在（0，2）上均匀分布

第七章 破产模型

学习目标

- ☐ 了解如下概念：盈余过程、总理赔过程、泊松过程、破产概率、最大总损失过程、调节系数、布朗运动以及净现金流入过程
- ☐ 了解寻找破产概率的四类方法
- ☐ 了解调节系数在确定最优再保险中的作用
- ☐ 熟悉连续时间破产概率的精确计算与近似计算
- ☐ 熟悉离散时间破产概率的计算
- ☐ 掌握并能运用各种方法计算破产概率

§7.1 盈余过程与破产概率

7.1.1 盈余过程

我们从影响保险人稳定经营最重要的方面——资产和负债入手。资产与负债的差额，一般称为**盈余**，简记做：

$$U(t)=A(t)-L(t), \qquad t\geqslant 0 \tag{7.1.1}$$

其中 $A(t)$ 表示时刻 t 的资产，$L(t)$ 表示时刻 t 的负债，$t=0$ 时刻的盈余被称为**初始盈余**、**初始资本**或**初始准备金**，简记为 u，即：$U(0)=u$。

现实中的 $A(t)$ 和 $L(t)$ 是非常多样化的，各个保险公司不同的保险经营模式和资产管理都会对应不同的资产负债价值过程。作为一个初步的理论模型，这里将对实际情况作较大的简化，这里的负债部分只考虑保险合同产生的理赔，资产部分则主要考虑保费收入，并且完全忽略对利率、投资收益率、通货膨胀、运营费用和保单红利等等因素的考虑。传统的风险理论对保费收入过程还有进一步的简化，假设保费收入按照固定的比例 c 线性增长，在现实中 c 为年保费收入。综上所述，传统的古典盈余过程模型为：

$$U(t)=u+ct-S(t), \qquad t\geqslant 0\ ;\ u\geqslant 0, c>0 \tag{7.1.2}$$

这是一个以 u 为初值、以时间为指标集的随机过程。

在模型（7.1.2）中，$\{S(t),\ t\geqslant 0\}$ 被称为**总理赔过程**，它表示从 0 到 t 时刻发生的所有理赔之和，因此，它是一个取值非负的以时间为指标集的随机过程，又简称为**理赔过程**。显然，模型（7.1.2）最核心的部分是

理赔过程。按照与第五章、第六章类似的研究思路，总理赔过程 $\{S(t), t\geqslant 0\}$ 由理赔次数和每次理赔额复合而成，所以有如下的总理赔过程表达：

$$S(t)=\begin{cases}X_1+X_2+\cdots+X_{N(t)}, & N(t)>0\\ 0, & N(t)=0\end{cases} \tag{7.1.3}$$

其中：

1. $N(t)$ 表示 $[0, t]$ 内的总理赔次数，取值非负整数，且 $N(0)=0$，称 $\{N(t), t\geqslant 0\}$ 为理赔次数过程（claim number process）；

2. 当 $N(t)=n>0$ 已知时，X_i，$i=1, 2, \cdots, n$ 表示 $[0, t]$ 内第 i 次理赔的金额；与第六章类似仍然称之为**理赔额变量**；

3. $S(t)$，$t\geqslant 0$ 表示 $[0, t]$ 内的总理赔额，且 $S(0)=0$。

由式（7.1.3）定义的总理赔过程 $\{S(t), t\geqslant 0\}$ 是对理赔的一种累积过程，在新的理赔发生时都会使总理赔额产生一个跳跃，而在第 i 次理赔与第 $i+1$ 次理赔之间，$S(t)$ 为常数。图7－1形象地说明了 $\{S(t), t\geqslant 0\}$ 过程的轨道特点。

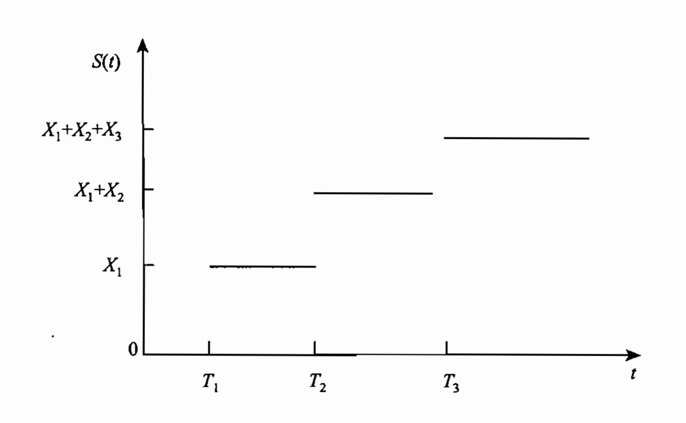


图7－1　总理赔过程 $\{S(t), t\geqslant 0\}$ 的轨道

7.1.2　破产概率

从盈余模型（7.1.2）看，一方面是连续不断的保费以速度 c 进行积累，另一方面则是不断地会有理赔需要支付，形成跳跃的支出，因此盈余过程也是一个跳跃的上升变化过程。在两次理赔之间，盈余按照比例 c 线性增长，一旦遇到理赔发生时，盈余立即（瞬时）按理赔额水平降低。而且，若某一时刻发生了数额很大的理赔，就很有可能马上出现盈余小于零的情况，通常人们形象并有点夸张地称这个事件为“破产”。

虽然在现实中发生这种 $U(t)\leqslant 0$ 的情况并不一定意味着保险公司马上会倒闭，可能也只是说明保险人需要及时追加资金来应付突然的保险责任，也许在考虑了其他许多因素后，保险人的财务状况并非很糟。但是，在简化的盈余模型中对这种破产事件进行研究无疑是非常有意义的，它对保险公司考虑财务预警系统以及对保险监管部门设计某些监管指标系统等问题有直接的参考指导作用。

以上关于破产的直观感觉可以用下面的一系列数学概念表述：

定义 7－1　破产时刻（ruin time）。称盈余过程（7.1.2）的如下随机变量

$$T=\inf\{t: t\geqslant 0, \ U(t)<0\} \tag{7.1.4}$$

为该盈余过程的破产时刻。实际上，T 为盈余首次出现负值的时刻。

定义 7－2　终极生存概率（survival probability）。称

$$\phi(u)=P(U(t)\geqslant 0, \forall t\geqslant 0) \tag{7.1.5}$$

为终极生存概率。

定义 7－3　终极破产概率（ruin probability）。称

$$\psi(u)=1-\varphi(u)=P(\exists t\geqslant 0, U(t)<0) \tag{7.1.6}$$

为终极破产概率。由终极破产概率的定义还可知 $\psi(u)=P(T<\infty)$。

由于式（7.1.6）中定义的终极破产概率没有时间上限，因此也称为**无限时间破产概率**。但在现实中由于保险公司往往关注的是在某一确定时期比如在 5 年、10 年或 20 年之内的经营状况，所以需要考虑有限时间内的破产概率。

定义 7－4　有限时间破产概率。称

$$\psi(u,t)=P(\exists s\in(0,t], U(s)<0) \tag{7.1.7}$$

为该盈余过程在时间 $(0, t]$ 内的破产概率，即盈余过程在有限时间内破产的概率。

由定义 7－4 可知有限时间破产概率 $\psi(u, t)=P(T\leqslant t)$，即 $\psi(u, t)=F_T(t)$。

破产概率 $\psi(u)$ 及 $\psi(u, t)$ 满足以下的基本性质，请读者自己练习验证。

性质 7－1　对于 $u_1\leqslant u_2$ 和 $0<t_1\leqslant t_2<\infty$，有以下结论成立：

（1）$\psi(u_2,t)\leqslant\psi(u_1,t), \forall t>0$；

（2）$\psi(u_2)\leqslant\psi(u_1)$；

（3）$\psi(u,t_1)\leqslant\psi(u,t_2)\leqslant\psi(u)$，

（4）$\lim\limits_{t\to\infty}\psi(u,t)=\psi(u)$

【例 7－1】　设某保单组合的理赔在 $0\leqslant t\leqslant 7$ 时间段中的记录如表 7－1 所示，又设保险人的初始盈余 $u=5$，保费收入比例 $c=4$。试描述该保单组合在这段时间内的盈余变化情况。

表 7－1　　某保单组合的理赔记录

理赔次序	1	2	3	4	5
发生时刻	0.5	2	2.75	4	6
理赔额	2.5	10	7	4	12

解：表 7－1 记录的是总理赔过程一条样本轨道的变化情况：

$$s(t)=\begin{cases}0, & 0\leqslant t<0.5\\ 2.5, & 0.5\leqslant t<2\\ 12.5, & 2\leqslant t<2.75\\ 19.5, & 2.75\leqslant t<4\\ 23.5, & 4\leqslant t<6\\ 35.5, & 6\leqslant t\leqslant 7\end{cases}$$

这里用 $s(t)$ 表明它是 $S(t)$ 的一条轨道，它描述了理赔过程（7.1.3）的一种情况。

对应于上述总理赔过程轨道的盈余过程的轨道为：

$$u(t)=5+4t-s(t)=\begin{cases}5+4t, & 0\leqslant t<0.5\\ 5+4t-2.5, & 0.5\leqslant t<2\\ 5+4t-12.5, & 2\leqslant t<2.75\\ 5+4t-19.5, & 2.75\leqslant t<4\\ 5+4t-23.5, & 4\leqslant t<6\\ 5+4t-35.5, & 6\leqslant t\leqslant 7\end{cases}$$

$$=\begin{cases}5+4t, & 0\leqslant t<0.5\\ 2.5+4t, & 0.5\leqslant t<2\\ -7.5+4t, & 2\leqslant t<2.75\\ -14.5+4t, & 2.75\leqslant t<4\\ -18.5+4t, & 4\leqslant t<6\\ -30.5+4t, & 6\leqslant t\leqslant 7\end{cases}$$

它描述了盈余过程（7.1.2）的一条轨道，其变化如图 7－2 所示。从图 7－2 中可以看出，在 $t=2.75$ 时，轨道第一次落到 t 轴之下，也就是说在 $t=2.75$ 时发生破产。

【例 7－2】 设保险人对于某类保单设置的初始盈余为 10 万元，估计年内将以每份 5 000 元的价格售出 100 张这类保单。假定每张保单是否发生理赔是相互独立的，总理赔费用占总保费 P 的 20%。又假定该类保单组合 n 年的总理赔 $S(n)$ 用均值为 $0.7P$、方差为 $2P$ 的正态分布近似，其中 P 表示年保费总收入。为了在每年末评估保险人的盈余状况，试计算该公司在第一年末出现负盈余的概率。

解：为简化计算，下面以 1 万元为一个货币单位，设初始盈余为 u，第一年的理赔额为 $S(1)$，依题意，有 $u=10$，$P=100\times 0.5$，则该公司在第一年末的盈余为：

$$\begin{aligned}U(1)&=\text{初始盈余}+\text{第一年的保费收入}-\text{第一年的理赔费用}-\text{第一年的理赔}\\&=u+P-0.2P-S(1)=10+100\times 0.5-0.2\times 100\times 0.5-S(1)\end{aligned}$$

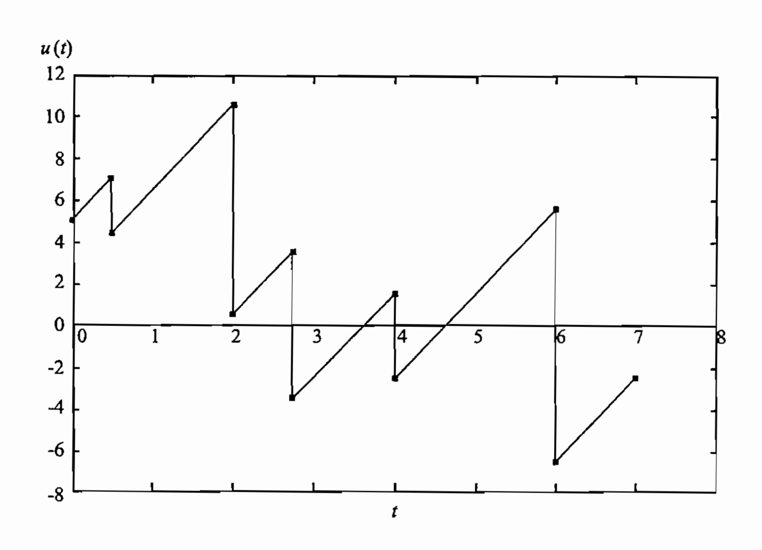


图 7－2　例 7－1 的盈余过程轨道示例 ■

$$=50-S(1)$$

经计算，有 $S(1)\sim N(35,10^2)$，因此，第一年年末出现负盈余的概率为：

$$P\{U(1)\leqslant 0\}=P\{S(1)\geqslant 50\}=1-\Phi\left(\frac{50-35}{10}\right)$$

$$=1-\Phi(1.5)=1-0.93319=0.067$$ ■

【例 7－3】 承上例，考虑第二年的情况。假定第二年内能够再售出 200 张同类保单，试计算第二年末出现负盈余的概率。

解： 这时，$P=(100+200)\times 0.5=150$，同样可得保险人在第二年末的盈余为：$U(2)=130-S(2)$。而 $S(2)\sim N(105,\ 17.3^2)$，第二年末出现负盈余的概率为：

$$P\{U(2)\leqslant 0\}=P\{S(2)\geqslant 130\}=1-\Phi\left(\frac{130-105}{17.3}\right)$$

$$=1-\Phi(1.443)=0.074$$ ■

从上面的两个例子可以看出，"破产"是一个非常多义的词汇，在例 7－3中只考虑了第二年末出现负盈余的情况，与例 7－2 的情形无关，这两个例子都只是计算了一个固定点的盈余情况。

正如"破产"这个词汇的字面含义所强调的，了解这个不利事件发生的概率具有特别重要的意义。因此，本章的主要任务之一就是要在各种条件下获得关于盈余过程破产和破产概率的性质。

§7.2　总理赔过程

通过上面的分析，我们知道盈余模型（7.1.2）最重要的部分是总理赔

过程（7.1.3），总理赔过程（7.1.3）又由理赔次数过程 $\{N(t), t\geq 0\}$ 和理赔额变量 X 复合而成。本节将分别对理赔次数过程和总理赔过程的基本性质进行讨论。

7.2.1　理赔次数过程——泊松过程

理赔次数过程是所谓计数随机过程的特例。

定义 7-5　随机过程 $\{N(t), t\geq 0\}$ 被称为**计数随机过程**，若

（1）取值为非负整数，且 $N(0)=0$；

（2）样本轨道为阶梯形，每次增加 1。

若计数过程是记录了某种事件发生的次数，那么由计数过程的定义知，在充分小的时间区间内，至多发生一次事件。为了刻画计数过程，有时可以考虑以下的三种描述方法：

方法一：对任意 $t\geq 0$，$h>0$，有 $P(N(t+h)>N(t)+1 \mid N(s), 0<s\leq t)=o(h)$

方法二：对任意 $t\geq 0$，考虑 $\lim\limits_{\Delta t\to 0+}\dfrac{1}{\Delta t}P[N(t+\Delta t)-N(t)=1 \mid N(s), 0<s\leq t]$，记该极限（若存在）为 $\lambda(t)$，称为计数过程的**强度函数**。则计数过程为在时间区间 $[t, t+\Delta t]$ 内，发生一次事件的概率为 $\lambda(t)\Delta t$，事件不发生的概率为 $1-\lambda(t)\Delta t$，发生两次以上事件的概率为 $o(\Delta t)$。

方法三：对任意 $t\geq 0$，当 $N(t)>0$ 时，记：

$$T_i=\inf\{t: t>T_{i-1}, N(t)>N(T_{i-1})\}, i=1,2,\cdots,N(t) \tag{7.2.1}$$

则有

$$0=T_0<T_1<T_2<\cdots<T_{N(t)}$$

依次表示 $[0, t]$ 上各次事件发生的时刻，且有

$$W_i=T_i-T_{i-1},\quad i=1,2,\cdots,N(t) \tag{7.2.2}$$

表示两次事件之间的时间间隔，一般称为**等待时间**（waiting time）变量，W_i 为非退化的取非负实值的随机变量。

图 7-3 是对理赔次数过程 $\{N(t), t\geq 0\}$ 的直观描述，其中 T_1，T_2，… 表示各次理赔发生的时刻。计数过程的轨道除 T_1，T_2，… 点外是连续的，而且在这些间断点右连续、左极限存在。

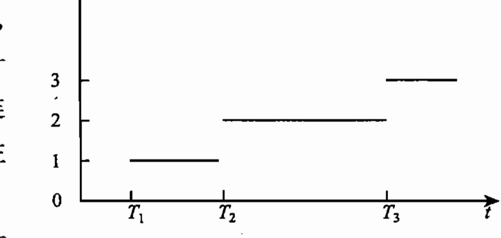


图 7-3　理赔次数过程 $\{N(t), t\geq 0\}$ 示例

最常见的计数过程是泊松（计数）过程，可以用定义 7-6、7-7 来定义泊松过程：

定义 7－6 计数过程 $\{N(t), t\geqslant 0\}$ 被称为强度参数为 λ 的泊松过程，若：

（1）独立增量，即对任意的 $0<s<t\leqslant v<u$，随机变量 $N(t)-N(s)$ 与 $N(u)-N(v)$ 独立；

（2）平稳增量，即对任意 $t\geqslant 0$，$h>0$，$N(t+h)-N(t)$ 的分布只与区间长度 h 有关；

（3）$N(t+h)-N(t)\sim \text{Poisson}(\lambda h)$。

定义 7－7 计数过程 $\{N(t), t\geqslant 0\}$ 被称为强度参数为 λ 的泊松过程，若：

（1）独立的平稳增量；

（2）对任意 $h>0$，有 $P(N(h)=1)=\lambda h+o(h)$；

（3）对任意 $h>0$，有 $P(N(h)\geqslant 2)=o(h)$。

独立增量意味着任意区间内的索赔次数与之前任意与其不相交的区间内的索赔次数互相独立，而平稳增量则说明固定区间内的索赔次数只依赖于区间长度，与起始位置无关，如不受时间趋势的影响。若一个随机过程同时满足平稳性和独立性，则独立的平稳增量过程就可以看做是从任一时刻开始的原过程。事实上，定义 7－6 中的条件（3）和定义 7－7 中的条件（2）、（3）已经暗含了平稳独立增量的性质。因此，平稳独立增量的假设只是为了更强调这一性质，从而使定义的阐述更加明确。另外，定义 7－6 中的条件（3）和定义 7－7 中的条件（2）、（3）是等价的，请读者自行证明。

下面我们不加证明地列出关于泊松过程的两个进一步的结论：

（1）对任意的 $t\geqslant 0$，当已知 $N(t)=n$ 的信息时，随机向量（T_1，T_2，…，T_n）与 n 个独立的且服从区间 $[0, t]$ 上均匀分布的随机变量形成的次序（按从小到大排序）统计量的随机向量同分布，即：

$$f_{T_1,\cdots,T_n}(t_1,\cdots,t_n\mid N(t)=n)=\frac{n!}{t^n},\qquad t_1\leqslant\cdots\leqslant t_n\leqslant t$$

（2）当 $T_{i-1}=t$ 及所有 $N(s)$，$s\leqslant t$ 的信息已知时，由式（7.2.2）定义的等待时间 W_i 为（条件）独立同分布的随机变量序列，且共同分布是均值为 $1/\lambda$ 的指数分布，其中 λ 为泊松过程的强度参数。

7.2.2 理赔总量过程——复合泊松过程

若聚合理赔过程（7.1.3）中的理赔次数过程 $\{N(t), t\geqslant 0\}$ 为泊松过程，则称式（7.1.3）为复合泊松过程。为强调起见，这里对复合泊松过程再次给出完整的定义。

定义 7－8 式（7.1.3）的总理赔过程 $\{S(t), t\geqslant 0\}$ 称为**复合泊松过**

程，若：

（1）计数过程 $\{N(t), t \geq 0\}$ 是强度参数为 λ 的泊松过程；

（2）$\{X_i, i=1, 2, \cdots\}$ 是独立同分布的随机变量序列，共同的分布函数记为 $F(x)$、密度函数记为 $f(x)$，X 的 k 阶原点矩（若存在）记为 p_k，$k=0, 1, 2, \cdots$；

（3）$\{N(t), t \geq 0\}$ 与 $\{X_i, i=1, 2, \cdots\}$ 独立。

复合泊松过程具有以下基本性质：

（1）平稳，独立增量，即对任意取定的 $t \geq 0$，$h>0$，$S(t+h)-S(t)$ 与

$$S(h)=\begin{cases} X_1+X_2+\cdots+X_{N(h)} & N(h)>0 \\ 0 & N(h)=0 \end{cases}$$

同分布，其共同分布为复合泊松分布：泊松参数为 λh，理赔额变量的分布函数为 $F(x)$。

（2）经过简单的计算，复合泊松过程的特征变量为：

$$E[S(t)]=\lambda t E(X)=\lambda t p_1 \tag{7.2.3}$$

$$Var[S(t)]=\lambda t E(X^2)=\lambda t p_2 \tag{7.2.4}$$

$$M_{S(t)}(r)=e^{\lambda t[M_X(r)-1]}\text{且属于 } M_X(r)\text{的定义域} \tag{7.2.5}$$

其中 $M_X(r)$ 为理赔额变量 X 的矩母函数。

（3）过程的轨道除 T_1，T_2，$\cdots$ 点外是连续的，而且在这些间断点右连续、左极限存在。当 T_i 已知时，$S(T_i)-S(T_i-0)$ 与 X 同分布。

在总理赔过程是复合泊松过程的基础上，可以进一步考虑相应的盈余过程，这也是传统的古典破产理论结论最多的盈余模型。

定义 7－9　称下面的过程 $\{U(t), t \geq 0\}$ 为（连续时间）**泊松盈余过程**（surplus process）：

$$U(t)=u+ct-S(t), \qquad t \geq 0$$

其中：

（1）u 为初始盈余，$u \geq 0$；

（2）$\{S(t), t \geq 0\}$ 为复合泊松过程，泊松参数 λ，理赔额变量 $X \sim F(x)$；

（3）c 为单位时间内收取的保费，满足 $c=(1+\theta)\lambda p_1$，$\theta>0$；

对于泊松盈余过程，有以下的基本结论：

$$E[U(t)]=E[u+ct-S(t)]=u+\theta\lambda p_1 t \tag{7.2.6}$$

$$Var[U(t)]=Var[S(t)]=\lambda p_2 t \tag{7.2.7}$$

$$M_{U(t)}(r)=e^{ur+ctr-\lambda t[M_X(r)-1]} \tag{7.2.8}$$

从式（7.2.6）和（7.2.7）可以看出泊松盈余过程的均值和方差都是 t 的线性函数。由于 $\theta>0$，则有

$$\lim_{t\to\infty} E[U(t)]=\lim_{t\to\infty}(u+\theta\lambda p_1 t)=+\infty \tag{7.2.9}$$

进而通过一定的推导，有

$$\lim_{u\to\infty}\psi(u)=0 \tag{7.2.10}$$

这表明：只要保费收入超过平均损失，保险人的长期利益就得到了充分的保障。但是在短期经营中，可能会因理赔的随机性出现大量理赔集中发生的情况，使得保险人的财务暂时出现赤字，发生破产。

§7.3 连续时间终极破产概率的计算

经过前面的准备，这一节我们开始讨论满足一定性质的盈余过程（7.1.2）的破产概率。这里的讨论都是基于定义 7－9 的泊松盈余过程进行的，对这个古典盈余过程破产概率的研究，主要围绕微积分方程、最大总损失和调节系数三个技术方法进行。

7.3.1 微分方程式与破产概率

本节将对复合泊松盈余过程导出关于破产概率 $\psi(u)$ 的微积分方程式，并对理赔额变量 X 的某种特殊情形求解方程，进而得到 $\psi(u)$ 的显式解。

定理 7－1 对于泊松盈余过程，破产概率 $\psi(u)$ 满足以下的微积分方程：

$$\psi'(u)=\frac{\lambda}{c}\psi(u)-\frac{\lambda}{c}\int_0^u\psi(u-x)dF(x)-\frac{\lambda}{c}[1-F(u)],u\geqslant 0 \tag{7.3.1}$$

$$\psi(u)=\frac{\lambda}{c}\int_0^u\psi(u-x)[1-F(x)]dx+\frac{\lambda}{c}\int_u^\infty[1-F(x)]dx,u\geqslant 0 \tag{7.3.2}$$

其中，$F(x)$ 为理赔额变量 X 的分布函数。特殊地，有

$$\psi(0)=\frac{1}{1+\theta} \tag{7.3.3}$$

证明： 首先证明式（7.3.1）。考虑在充分小的时间（0，Δt］内至多发生一次理赔，所以，下面按照是否发生理赔两种情况讨论：

1. 在（0，Δt］内没有理赔，概率为 $1-\lambda\Delta t$。在时刻 Δt 的盈余为 $u+c\Delta t$，而且由泊松过程的独立增量性，这时的破产概率为 $\psi(u+c\Delta t)$。

2. 若（0，Δt］内发生一次理赔，概率为 $\lambda\Delta t$，若理赔额为 x，则发生该理赔额的瞬间概率为 $dF(x)$。可进一步讨论在时刻 Δt 之前破产或不破产两种情形：

（1）在（0，Δt］内未破产，即：$0\leqslant x\leqslant u+c\Delta t$。在时刻 Δt 的盈余为 $u+c\Delta t-x$，而且由泊松过程的独立增量性，这时的破产概率为 $\psi(u+c\Delta t-x)$，当 x 取值于整个（0，$u+c\Delta t$］时，则

$$\int_0^{u+c\Delta t}\psi(u+c\Delta t-x)dF(x)$$

表示破产概率的累加结果。

(2) 在 $(0, \Delta t]$ 内破产，即：$x > u + c\Delta t$。则 $1 - F(u + c\Delta t)$ 表示所有这类事件的总概率。

总结上面的两种情况，有

$$\psi(u) = (1 - \lambda\Delta t)\psi(u + c\Delta t) + \lambda\Delta t\left[\int_0^{u+c\Delta t}\psi(u + c\Delta t - x)dF(x)\right.$$
$$\left. + 1 - F(u + c\Delta t)\right]$$

整理上式，得：

$$\frac{\psi(u + c\Delta t) - \psi(u)}{c\Delta t} = \frac{\lambda}{c}\psi(u + c\Delta t) - \frac{\lambda}{c}\int_0^{u+c\Delta t}\psi(u + c\Delta t - x)dF(x)$$
$$- \frac{\lambda}{c}[1 - F(u + c\Delta t)]$$

令 $\Delta t \to 0$，则 $\psi'(u)$存在且式（7.3.1）得证。

下面证明式（7.3.2）。将式（7.3.1）两边积分，有

$$\int_0^u \psi'(v)dv = \frac{\lambda}{c}\int_0^u \psi(v)dv - \frac{\lambda}{c}\int_0^u\int_0^v \psi(v - x)dF(x)dv$$
$$- \frac{\lambda}{c}\int_0^u [1 - F(x)]dx$$

即：

$$\psi(u) - \psi(0) = \frac{\lambda}{c}\int_0^u \psi(v)dv - \frac{\lambda}{c}\int_0^u\int_0^{u-x}\psi(v)dvdF(x) - \frac{\lambda}{c}\int_0^u [1 - F(x)]dx$$
$$= \frac{\lambda}{c}\left\{\int_0^u \psi(v)dv - \left[\int_0^{u-x}\psi(v)dvF(x)\Big|_{x=0}^{u} + \int_0^u \psi(u - x)F(x)dx\right]\right.$$
$$\left. - \int_0^u [1 - F(x)]dx\right\}$$
$$= \frac{\lambda}{c}\left\{\int_0^u \psi(u - x)dx - \int_0^u \psi(u - x)F(x)dx - \int_0^u [1 - F(x)]dx\right\}$$
$$= \frac{\lambda}{c}\left\{\int_0^u \psi(u - x)[1 - F(x)]dx - \int_0^u [1 - F(x)]dx\right\}$$

上述推导过程采用了分部积分方法和已知条件 $F(0) = 0$。即：

$$\psi(u) = \psi(0) + \frac{\lambda}{c}\left\{\int_0^u \psi(u - x)[1 - F(x)]dx\right.$$
$$\left. - \int_0^u [1 - F(x)]dx\right\} \tag{7.3.4}$$

令 $u \to \infty$，由式（7.3.4）有

$$0 = \psi(0) + \frac{\lambda}{c}\left\{\lim_{u\to\infty}\left(\int_0^u \psi(u - x)[1 - F(x)]dx\right)\right.$$
$$\left. - \int_0^\infty [1 - F(x)]dx\right\}$$

另外，可以证明：

$$\lim_{u\to\infty}\left(\int_0^u \psi(x)dx\right) = \lim_{u\to\infty}\left(\int_0^u \psi(u-x)F(x)dx\right)$$

则有

$$\psi(0) = \frac{\lambda}{c}\int_0^\infty [1-F(x)]dx = \frac{\lambda}{c}p_1 = \frac{1}{(1+\theta)}$$

由此，式（7.3.3）得证，再将上式代入式（7.3.4）可得式（7.3.2）的结论。 ■

虽然定理 7－1 得到了破产概率的微积分表达式，但要获得式（7.3.1）的解 $\psi(u)$ 的解析表达式并不容易，只有一些特殊的情形才会得到这种表达式。

定理 7－1 的式（7.3.3）具有特别的含义，即在初始资金为零的情况下，所收取的保费不够支付理赔的概率。从式（7.3.3）可看出，这个概率仅与附加保费率有关，与理赔次数过程的参数 λ 和理赔额的分布 $F(x)$ 无关。

定理 7－2 设在泊松盈余过程中，理赔额变量服从均值为 $1/\alpha$ 的指数分布，则破产概率为：

$$\psi(u) = \frac{1}{1+\theta}\exp\left(-\frac{\alpha\theta}{1+\theta}u\right), \qquad u\geqslant 0 \tag{7.3.5}$$

证明：将 $F(x)=1-e^{-\alpha x}$ 代入式（7.3.1），可得：

$$\psi'(u) = \frac{\lambda}{c}\psi(u) - \frac{\lambda\alpha}{c}e^{-\alpha u}\int_0^u e^{\alpha x}\psi(x)dx - \frac{\lambda}{c}e^{-\alpha u}, u\geqslant 0 \tag{7.3.6}$$

方程（7.3.6）两边求导，有

$$\psi''(u) = \frac{\lambda}{c}\psi'(u) + \frac{\lambda\alpha^2}{c}e^{-\alpha u}\int_0^u e^{\alpha x}\psi(x)dx - \frac{\lambda\alpha}{c}\psi(u) + \frac{\lambda\alpha}{c}e^{-\alpha u}, u\geqslant 0 \tag{7.3.7}$$

将（7.3.6）两边同乘 α 后与（7.3.7）相加，得：

$$\psi''(u) + \left(\alpha - \frac{\lambda}{c}\right)\psi'(u) = 0, \qquad u\geqslant 0 \tag{7.3.8}$$

另由 $c=(1+\theta)\lambda p_1$，$\theta>0$，则

$$\left(\alpha - \frac{\lambda}{c}\right) = \frac{\alpha\theta}{1+\theta}$$

所以，微分方程（7.3.8）的通解为：

$$\psi(u) = k_1 + k_2 e^{-\frac{\alpha\theta}{1+\theta}u}, \qquad u\geqslant 0$$

其中 k_1 和 k_2 为任意常数。为了确定这两个任意常数，分别令 $u=0$ 和 $u\to\infty$，得 $k_1=0$ 和 $k_2=\psi(0)$，因此有：

$$\psi(u) = \frac{1}{1+\theta}e^{-\frac{\alpha\theta}{1+\theta}u}, \qquad u\geqslant 0$$ ■

定理 7－3 对于泊松盈余过程，当盈余首次降至初始盈余 u 以下时，

其当时盈余的负值（一般称为亏损变量 deficit）的概率密度为：

$$P(u+y<-U(T)<u+y+dy \mid T<\infty)=\frac{1-F(y)}{(1+\theta)p_1}dy, \qquad y\geqslant 0$$

(7.3.9)

其中 $F(y)$ 是理赔额变量的分布函数。

该定理的证明过程与定理 7－1 的证明方法类似，但是有一定的难度，这里略去其证明，读者可查阅参考文献（Bower etc，1997）。

定理 7－3 的含义同样在于只考虑所收取的保费不够支付理赔的概率，而不关心初始的盈余水平，因为对于某种具体的保险业务来说，开始很难确定究竟要给它分配多少初始资本。所以，不考虑初始盈余，单单分析保费不够支付理赔的概率也具有很重要的意义。

由于 $\int_0^{\infty}[1-F(y)]dy=p_1$，所以，

$$f_{L_1}(y)=\frac{1-F(y)}{p_1}, \qquad y\geqslant 0 \tag{7.3.10}$$

可看做是某个随机变量的概率密度，我们一般用 L_1 表示这类随机变量。它相当于是对随机变量 X 的分布进行的变换，这种变换在风险理论讨论中很有意义。由式（7.3.10）的定义，经过简单的计算可以得到 L_1 的矩母函数（若存在）的表达式：

$$M_{L_1}(r)=\frac{1}{rp_1}(M_X(r)-1), \qquad r\geqslant 0 \tag{7.3.11}$$

7.3.2　最大总损失与破产概率

当初始盈余为 u 时，盈余过程（7.1.2）的终极生存概率为：

$$\begin{aligned}\phi(u)&=P(T=\infty)=P(U(t)\geqslant 0,\ \forall t\geqslant 0)\\&=P(S(t)-ct\leqslant u,\forall t\geqslant 0)=P(\max_{t\geqslant 0}[S(t)-ct]\leqslant u)\end{aligned}$$

定义 7－10　盈余过程 $\{U(t),\ t\geqslant 0\}$ 的最大总损失（maximal aggregate loss）随机变量为：

$$L=\max_{t\geqslant 0}[S(t)-ct]$$

当 $\theta>0$ 时，有

$$\lim_{t\to\infty}(S(t)-ct)=-\infty$$

自然有 $L\geqslant 0$，而且对任意的 $u\geqslant 0$，有

$$\psi(u)=1-P(L\leqslant u)=1-F_L(u)>0$$

利用最大总损失变量的上述定义，我们可以得到关于这个变量的一些性质和结论。

定理 7－4　若 $\{S(t),\ t\geqslant 0\}$ 为复合泊松过程，则有以下结论成立：

(1) $$\psi(u)=1-\sum_{n=0}^{\infty}\left(\frac{1}{1+\theta}\right)^n\left(\frac{\theta}{1+\theta}\right)F_{L_1}^{*n}(u) \tag{7.3.12}$$

(2) $$M_L(r)=\frac{\theta r p_1}{1+(1+\theta)rp_1-M_X(r)}$$
$$=\frac{\theta}{1+\theta}+\frac{1}{1+\theta}\cdot\frac{\theta[M_X(r)-1]}{1+(1+\theta)rp_1-M_X(r)},\qquad r\geqslant 0 \qquad (7.3.13)$$

其中 L_1 的定义见（7.3.10）。

证明：考虑如下的随机序列：

（1）$S(t)-ct$ 的第 $k(\ k\geqslant 0\ )$ 次记录产生的时刻为：

$$\widetilde{T}_k=\inf_{t\geqslant \widetilde{T}_{k-1}}\{t;S(t)-ct>\widetilde{L}_{k-1}\},\qquad \widetilde{T}_0=0$$

$\widetilde{T}_k$ 相当于以 $\widetilde{L}_{k-1}$ 为初始盈余的盈余过程首次低于 $\widetilde{L}_{k-1}$ 的时刻。

（2）$S(t)-ct$ 第 $k(\ k\geqslant 0\ )$ 次记录的值为：

$$\widetilde{L}_k=S(\widetilde{T}_k)-c\widetilde{T}_k,\ \widetilde{L}_0=0$$

利用泊松过程的性质可以证明：最大总损失 L 可以表示为：

$$L=L_1+L_2+\cdots+L_N$$

其中：

① N 表示 $S(t)-ct$ 在 $t\in[0,\infty]$的轨道上破记录（“破产”）的总数，则

$$P(N=n)=P(n\text{ 次“破产”,一次不“破产”})=\psi^n(0)[1-\psi(0)]$$
$$=\left(\frac{1}{1+\theta}\right)^n\left(\frac{\theta}{1+\theta}\right)$$

N 服从参数为 $P=\dfrac{\theta}{1+\theta}$的几何分布。

② L_n表示两次记录的差值，也就是前面讨论过的亏损变量，所以随机变量列 $\{L_n=\widetilde{L}_n-\widetilde{L}_{n-1},\ n=1,\ 2,\ \cdots\}$ 独立同分布，共同的分布为 $f_{L_1}(x)$。

然后按照复合分布的定义自然有：

$$\psi(u)=1-F_L(u)=1-\sum_{n=0}^{\infty}P(N=n)F_{L_1}^{*n}(u)$$
$$=1-\sum_{n=0}^{\infty}\left(\frac{1}{1+\theta}\right)^n\left(\frac{\theta}{1+\theta}\right)F_{L_1}^{*n}(u)$$

$$M_L(r)=M_N(\log M_{L_1}(r))=\frac{\theta}{1+\theta-M_{L_1}(r)}$$

将式（7.3.11）代入，有

$$M_L(r)=\frac{\theta r p_1}{1+(1+\theta)rp_1-M_X(r)},\qquad r\geqslant 0$$

经过适当的整理，上式与式（7.3.13）相同。

式（7.3.13）表明最大总损失 L 服从如下的混合分布：

$$F_L(0)=\frac{\theta}{1+\theta}$$

$L>0$ 部分的矩母函数为：$\dfrac{\theta[M_X(r)-1]}{1+(1+\theta)rp_1-M_X(r)}$。

【例 7-4】 若 $X\sim$ Pareto（α，β），$\alpha>1$，$\beta>0$，当 $\theta=0.2$；$\alpha=3$，$\beta=1\ 000$ 时，计算破产概率。

解： 由已知

$$F(x)=1-\left(\frac{\beta}{x+\beta}\right)^{\alpha},\qquad x>0$$

$$p_1=\frac{\beta}{\alpha-1}=500$$

则有

$$f_{L_1}(x)=\frac{1}{p_1}\left(\frac{\beta}{x+\beta}\right)^{\alpha}$$

所以，$L_1\sim$ Pareto（$\alpha-1$，β）。我们对初始盈余取不同的值，再利用式（7.3.12），可以计算破产概率的数值结果（见表 7-2）。

表 7-2　　不同初始盈余下的破产概率

u	100	500	1 000	5 000	10 000	25 000
$\frac{u}{p_1}$	0.2	1	2	10	20	50
$1-\psi(u)$	0.193	0.276	0.355	0.687	0.852	0.975
$\psi(u)$	0.807	0.724	0.645	0.313	0.148	0.025

■

推论 7-1 泊松盈余过程 $\{U(t),\ t\geqslant 0\}$ 的破产概率满足：

$$-\int_0^{\infty}e^{ru}d\psi(u)=\frac{1}{1+\theta}\cdot\frac{\theta[M_X(r)-1]}{1+(1+\theta)rp_1-M_X(r)},\qquad r\in(0,\gamma)\tag{7.3.14}$$

式（7.3.14）右边为 $1-\psi(u)$ 对应的矩母函数。

下面我们来看推论 7-1 的一个应用。

若 X 为指数分布的线性组合，例如：$f(x)=c_1\beta_1e^{-\beta_1x}+c_2\beta_2e^{-\beta_2x},\beta_1>\beta_2>0$，$c_1>0,c_2>0,c_1+c_2=1$。则

$$p_1=c_1\frac{1}{\beta_1}+c_2\frac{1}{\beta_2}$$

$$M_X(r)=c_1\frac{\beta_1}{\beta_1-r}+c_2\frac{\beta_2}{\beta_2-r},\qquad \beta_1>r>0,r\neq\beta_2\tag{7.3.15}$$

式（7.3.15）的左边作为矩母函数，按照期望存在的情况其定义域范围是 $r<\gamma=\beta_2=\min\{\beta_1,\ \beta_2\}$。然而，式（7.3.15）的右边作为一般的函数，其定义域可以扩展到 $r>0$ 且 $r\neq\beta_1$ 和 $r\neq\beta_2$ 的一切实数。为简单起见，我们仍然用符号 $M_X(r)$ 表示这个扩展了定义域后的函数。

将上述计算结果代入式（7.3.14），有：

$$-\int_0^\infty e^{ru}d\psi(u)=\frac{\theta}{1+\theta}\cdot\frac{g_1(r)}{g_2(r)}$$

其中：$g_1(r)$为r的一次多项式，$g_2(r)$为r的二次多项式。最终可表示为：

$$-\int_0^\infty e^{ru}d\psi(u)=k_1\frac{\alpha_1}{\alpha_1-r}+k_2\frac{\alpha_2}{\alpha_2-r},\qquad \alpha_1>0,\alpha_2>0$$

又由$\psi(\infty)=0$，则有

$$\psi(u)=k_1e^{-\alpha_1u}+k_2e^{-\alpha_2u}$$

【例7-5】 试由式（7.3.14）推出定理7-2的结论。

解：X的矩母函数为：

$$M_X(r)=\frac{\alpha}{\alpha-r}=\frac{1}{1-p_1r},\qquad r\in(0,\gamma)$$

代入（7.3.13）式，有

$$-\int_0^\infty e^{ru}d\psi(u)=\frac{1}{1+\theta}\left[\frac{\theta}{\theta-(1+\theta)p_1r}\right]$$

令$c_1=\dfrac{1}{1+\theta}$及$r_1=\dfrac{\theta}{(1+\theta)p_1}$，则有

$$\int_0^\infty e^{ru}d[1-\psi(u)]=c_1\frac{r_1}{r_1-r}$$

上式的右边是某个指数分布函数的矩母函数，$\psi(u)$的唯一解是$\psi(u)=c_1e^{-r_1u}$。 ■

§7.4 破产概率与调节系数

前面几节中讨论了与盈余过程有关的几个方面，并从不同的角度来看待“破产”这个特殊事件的概率。在第一节中我们曾强调过，本章的目的就是要获得对破产概率的具体表达。从盈余过程的数学模型来看，破产概率当然与保险人的初始盈余u直接相关。另一个影响破产概率的主要因素显然是保险费率c，而c又直接与理赔次数过程和理赔额变量有关，所以，希望得到破产概率与上述参数的关系。

要想直接得到函数表达关系非常困难，Lundberg（1919）发现了一个间接的表达方法，即引入一个能起到中介作用的参数，称为**Lundberg系数**或**调节系数**。

7.4.1 调节系数的概念与性质

定义7-11 对泊松盈余过程，称满足以下方程：

$$\lambda+cr=\lambda M_X(r),\qquad r\in(0,\gamma)\tag{7.4.1}$$

的非零正解（记做R）为该过程的调节系数。

方程（7.4.1）是关于自变量 r 的一个隐式方程，方程左边是关于 r 的一个线性函数，方程右边是理赔额 X 的矩母函数，是一个下凸函数，两者的关系见图 7－4。

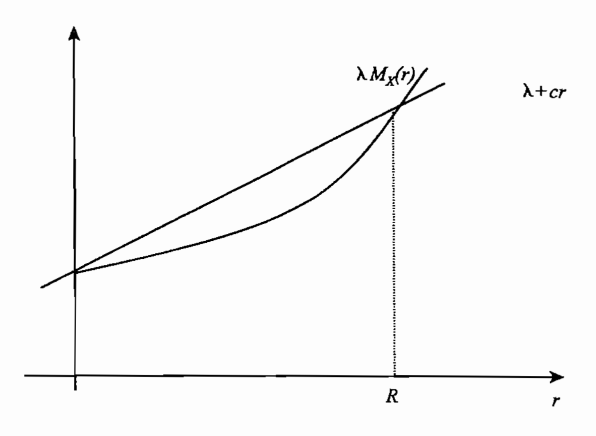


图 7－4　调节系数方程示意图

下面的定理给出了调节系数 R 的性质。

定理 7－5　设泊松盈余过程中理赔额变量 X 的矩母函数的定义域为 $(0,\ \gamma)$，其中 $\gamma \leq +\infty$。在 $c > \lambda p_1$ 的假设下，方程（7.4.1）有唯一的正根 $R \in (0,\ \gamma)$。

证明： 令 $g(r) = \lambda M_X(r) - \lambda - cr, r \in (0,\gamma)$，若 $g(r)$ 具有以下性质：

(1) $g(0) = 0$；(2) $g'(0) < 0$；(3) $\lim\limits_{r\to\gamma} g(r) = +\infty$；(4) $g''(r) > 0$，$\forall r \in (0,\ \gamma)$。

则方程 $g(r) = \lambda M_X(r) - \lambda - cr$ 有唯一正根。

事实上，

$$g(0) = \lambda M_X(0) - \lambda = \lambda - \lambda = 0$$

$$g'(0) = M_X'(0) - c = \lambda p_1 - c < 0$$

若 $\gamma < +\infty$，由 $M_X(r)$ 的定义知 $M_X(\gamma) = +\infty$，从而有

$$\lim_{r\to\gamma} g(r) = +\infty$$

若 $\gamma = +\infty$，由 $M_X(r) = 1 + p_1 r + \frac{p_2}{2}r^2 + \cdots > \frac{p_2}{2}r^2$，有

$$\lim_{r\to\infty} g(r) = \lim_{r\to\infty}(\lambda M_X(r) - \lambda - cr) > \lim_{r\to\infty}\left(\lambda \frac{p_2}{2}r^2 - \lambda - cr\right) = +\infty$$

最后，有

$$g''(r) = \lambda M''_X(r) = \lambda E(X^2 e^{rX}) > 0$$

即 $g(r)$ 为下凸函数，结论得证。 ■

若将 $c = (1+\theta)\lambda p_1$ 代入（7.4.1），可得：

$$1 + (1+\theta)p_1 r = M_X(r) \qquad (7.4.2)$$

这个调节系数方程与泊松盈余过程的泊松参数无关，在使用上有其方便之处。

方程(7.4.1)和(7.4.2)往往很难得到调节系数的显式解，这也正是 Lundberg 引入调节系数这个中介参数的原因。在不能获得 R 的精确值时，对 R 的可能取值范围作出估计也很有意义。

定理 7－6　调节系数 R 满足不等式：

$$R < \frac{2\left(\frac{c}{\lambda} - p_1\right)}{p_2} \tag{7.4.3}$$

证明：

$$\lambda + cR = \lambda M_X(R)$$

$$= \lambda\left[1 + p_1 R + \frac{p_2}{2}R^2 + \frac{p_3}{3}R^3 + \cdots\right] > \lambda\left[1 + p_1 R + \frac{p_2}{2}R^2\right]$$

$$= \lambda + \lambda p_1 R + \frac{\lambda p_2}{2}R^2$$

$$\Rightarrow (c - \lambda p_1)R > \frac{\lambda p_2}{2}R^2$$

$$\Rightarrow R < \frac{2[c - \lambda p_1]}{\lambda p_2} = \frac{2\theta p_1}{p_2} \tag{7.4.4}$$

定理 7－7　若理赔额 X 有上界 M，则有

$$R > \frac{1}{M}\log\frac{c}{\lambda p_1} = \frac{1}{M}\log(1+\theta) \tag{7.4.5}$$

证明： 为证式（7.4.5），先证以下不等式：

$$e^{Rx} \leqslant \frac{x}{M}e^{RM} + 1 - \frac{x}{M}, \qquad \forall\, 0 \leqslant x \leqslant M \tag{7.4.6}$$

事实上，从 e^x 的泰勒展开式知，对于任何 $0 \leqslant x \leqslant M$，有

$$\frac{x}{M}e^{RM} + 1 - \frac{x}{M} = \frac{x}{M}\sum_{k=0}^{\infty}\frac{(RM)^k}{k!} + 1 - \frac{x}{M}$$

$$= 1 + \sum_{k=1}^{\infty}\frac{R^k M^{k-1} x}{k!} \geqslant 1 + \sum_{k=1}^{\infty}\frac{(Rx)^k}{k!} = e^{Rx}$$

根据式（7.4.1），

$$\lambda + cR = \lambda M_X(R) = \lambda\int_0^M e^{Rx}f(x)\,dx \leqslant \lambda\int_0^{\infty}\left(\frac{x}{M}e^{RM} + 1 - \frac{x}{M}\right)f(x)\,dx$$

$$= \frac{\lambda p_1}{M}e^{RM} + \lambda - \frac{\lambda p_1}{M}$$

因此有

$$\frac{c}{\lambda p_1} \leqslant \frac{1}{RM}[e^{RM} - 1] = 1 + \frac{RM}{2} + \frac{(RM)^2}{3!} + \cdots < e^{RM}$$

从而结论得证。 ■

【例 7－6】　设泊松盈余过程的理赔额 $X \sim \begin{pmatrix} 10\ 000 & 25\ 000 \\ 0.9 & 0.1 \end{pmatrix}$，附加费率 $\theta = 0.2$。试写出调节系数的方程并估计 R。

解： 由已知

$$p_1 = 0.9 \times 10\ 000 + 0.1 \times 25\ 000 = 11\ 500$$

$$M_X(r) = 0.9e^{10\ 000r} + 0.1e^{25\ 000r}$$

又知 $\theta=0.2$，代入式（7.4.2）得：

$$1+13\ 800r=0.9e^{10\ 000r}+0.1e^{25\ 000r}$$

从该方程经过计算机迭代可解出调节系数 R 的近似解为 2.6×10^{-5}。另外，由于：

$$p_2=0.9\times10\ 000^2+0.1\times25\ 000^2=1.525\times10^8$$

按 R 的上界估计式（7.4.3），有

$$R<\frac{2\left(\frac{c}{\lambda}-p_1\right)}{p_2}=\frac{2(13\ 800-11\ 500)}{152\ 500\ 000}=3.02\times10^{-5}$$

所以，在这个例子中 R 的上界估计式（7.4.3）还是比较紧的。 ■

在一些特殊情况下，比如当理赔额变量服从指数分布或者退化为常数时，调节系数有比较简单的表达。

【例 7-7】 设复合泊松分布中的理赔额变量服从均值为 $1/\alpha$ 的指数分布，试求相应的调节系数。

解： 由于 $M_X(r)=\frac{\alpha}{\alpha-r}$，调节系数方程（7.4.2）为：

$$1+\frac{(1+\theta)r}{\alpha}=\frac{\alpha}{\alpha-r}$$

经过简单的计算，该方程的正根即调节系数为：

$$R=\frac{\theta}{1+\theta}\alpha$$

【例 7-8】 设理赔额变量退化为常数 1，试求相应的调节系数。

解： 由于 $M_X(r)=E(e^r)=e^r$，调节系数方程（7.4.2）应为：

$$1+(1+\theta)r=e^r$$

即求解直线 $y=1+(1+\theta)r$ 与标准指数函数曲线 $y=e^r$ 的交点，对于不同的 θ 值，可用数值方法求解上述方程获得调节系数，表 7-3 列举几个计算结果。

表 7-3　　不同 θ 值下的调节系数 R

θ	0.2	0.4	0.6	0.8	1.0	1.2
R	0.767	0.850	0.945	1.043	1.146	1.254

■

7.4.2 用调节系数表达破产概率

实际上，破产概率可以用调节系数 R 的函数来表示，最一般的表达式由以下定理给出。

定理 7-8 对于泊松盈余过程，若调节系数方程（7.4.2）的解 R 存

在，则破产概率 $\psi(u)$ 可以表示为：

$$\psi(u)=\frac{e^{-Ru}}{E[e^{-RU(T)}\mid T<\infty]},u\geqslant 0 \tag{7.4.7}$$

证明：对于任意的 $t>0$，$r>0$，考虑：

$$E[e^{-rU(t)}]=E[e^{-rU(t)}\mid T\leqslant t]P\{T\leqslant t\}+E[e^{-rU(t)}\mid T>t]P\{T>t\} \tag{7.4.8}$$

因为 $U(t)=u+ct-S(t)$，方程的左边等于：

$$\exp\{-ru-rct+\lambda t[M_X(r)-1]\}$$

右边第一项中的 $U(t)$ 在 $T\leqslant t$ 的条件下可变形为：

$$U(t)=U(T)+U(t)-U(T)=U(T)+c(t-T)-[S(t)-S(T)],T\leqslant t$$

当 T 给定时，$[S(t)-S(T)]$ 与 $U(T)$ 独立，而且前者为以 $t-T$ 为指标集的复合泊松过程，任意 t 时刻的泊松参数为 $\lambda(t-T)$，后者为确定的值。则式（7.4.8）的右边第一项可写成：

$$E[\exp(-rU(T))\exp\{-rc(t-T)+\lambda(t-T)[M_X(r)-1]\}\mid T\leqslant t]P\{T\leqslant t\}$$

若 R 为方程 $-rc+\lambda[M_X(r)-1]=0$ 的解，将 R 代入方程（7.4.8）两边，进一步简化为：

$$e^{-Ru}=E[e^{-RU(T)}\mid T\leqslant t]P(T\leqslant t)+E[e^{-RU(t)}\mid T>t]P(T>t) \tag{7.4.9}$$

当 $t\to\infty$ 时，式（7.4.9）右边第一项将收敛于 $E[e^{-RU(T)}\mid T<\infty]\ \psi(u)$。

现在只要证明式（7.4.9）右边第二项当 $t\to\infty$ 时趋于 0，定理就得到了证明。

为此，令 $\alpha=c-\lambda p_1$，$\beta^2=\lambda p_2$，可得：

$$E[U(t)]=E[u+ct-S(t)]=u+\alpha t,Var[U(t)]=Var[S(t)]=\beta^2 t$$

由于$\lim\limits_{t\to\infty}(u+\alpha t-\beta t^{\frac{2}{3}})=+\infty$，因此对充分大的 t，可以将式（7.4.9）右边第二项再展开，有

$$\begin{aligned}&E[e^{-RU(t)}\mid T>t]P\{T>t\}\\&=E[e^{-RU(t)}\mid T>t,0\leqslant U(t)\leqslant u+\alpha t-\beta t^{\frac{2}{3}}]P[T>t,0\leqslant U(t)\leqslant u\\&\quad+\alpha t-\beta t^{\frac{2}{3}}]+E[e^{-RU(t)}\mid T>t,U(t)>u+\alpha t-\beta t^{\frac{2}{3}}]\\&\quad\cdot P[T>t,U(t)>u+\alpha t-\beta t^{\frac{2}{3}}]\end{aligned}$$

而当 T 给定时，在 $T>t$ 的部分，有 $U(t)>0$，所以 $E[e^{-RU(t)}\mid T>t]\leqslant 1$，从而有

$$\begin{aligned}&E[e^{-RU(t)}\mid T>t,0\leqslant U(t)\leqslant u+\alpha t-\beta t^{\frac{2}{3}}]P\{T>t,0\leqslant U(t)\leqslant u+\alpha t-\beta t^{\frac{2}{3}}\}\\&\leqslant P\{U(t)\leqslant u+\alpha t-\beta t^{\frac{2}{3}}\}\leqslant t^{-\frac{1}{3}}\end{aligned}$$

上面的最后一个不等式由概率论中的切比雪夫不等式得到。同时有：

$$E[e^{-RU(t)}\mid T>t,U(t)>u+\alpha t-\beta t^{\frac{2}{3}}]P\{T>t,U(t)>u+\alpha t-\beta t^{\frac{2}{3}}\}$$

$$\leqslant \exp[-R(u+\alpha t-\beta t^{\frac{2}{3}})]$$

所以，当 $t \to \infty$ 时，式（7.4.9）右边趋于0，定理得证。■

定理7-8是关于破产概率的一个精确表达式，通过它还可以获得破产概率的近似表达。容易看出，当 $T<\infty$ 时，有 $U(T)<0$，所以 $E[e^{-RU(T)} \mid T<\infty]>1$，因此有

$$\psi(u) \leqslant e^{-Ru}, u \geqslant 0 \tag{7.4.10}$$

式（7.4.10）给出了破产概率的上界，具有一定的可操作性，被广泛使用，先看一个典型的例子。

【例7-9】 由定理7-8推导定理7-2关于指数分布破产概率的结论。

解： 考虑均值为 $\frac{1}{\alpha}$ 的指数理赔额分布，由例7-7可得 $R=\frac{\theta}{1+\theta}\alpha$。下面考虑 $-U(T)$ 的分布。假如破产发生在时刻 T，令 $\tilde{u}$ 表示破产时刻前一个瞬间的盈余（$U(T-)$），因为下一个瞬间时刻破产发生，所以破产时刻 T 发生的理赔额必然大于 $\tilde{u}$；对于任意的 $y>0$，若 $-U(T)>y$ 发生，则破产时刻的理赔额必大于 $\tilde{u}+y$，因此事件 $-U(T)>y$ 的条件概率为：

$$P(-U(T)>y \mid U(T-)=\tilde{u}, T<\infty)$$
$$=P(X>\tilde{u}+y \mid X>\tilde{u})$$
$$=\frac{\alpha\int_{\tilde{u}+y}^{\infty} e^{-\alpha x}dx}{\alpha\int_{\tilde{u}}^{\infty} e^{-\alpha x}dx}=e^{-\alpha y}$$

所以，条件随机变量 $[-U(T) \mid T<\infty]$ 仍然服从均值为 $\frac{1}{\alpha}$ 的指数分布，即：

$$E[e^{-RU(T)} \mid T<\infty]=\frac{\alpha}{\alpha-R}=\frac{1}{1+\theta}$$

因此有

$$\psi(u)=\frac{1}{1+\theta}e^{-\frac{\alpha\theta}{1+\theta}u}=\frac{1}{1+\theta}e^{-\frac{\theta}{1+\theta}\cdot\frac{u}{p_1}}$$ ■

显然，这个结论与定理7-2的结果一致。

§7.5　离散时间破产模型

我们在前面四节学习了连续时间破产模型并对终极破产概率的计算从三个角度进行了研究。连续时间模型中，我们需要持续不断地检查公司的盈余，但在现实中往往很难实现真正的连续观测，因为时间都是离散的，

例如：年、季、月等，更可能采取的方法是定期检查盈余。为此，我们需要考虑离散时间的盈余过程。

定义 7－12 称下面的过程$\{U(n),n=0,1,\cdots\}$为**盈余序列**：

$$U(n)=u+cn-S(n),S(0)=0 \tag{7.5.1}$$

其中：

（1）u 为初始盈余，$u\geqslant0$；

（2）c 为单位时间（一般为一年）内的平均保费收入，一般要求 $cn=(1+\theta)E[S(n)]>E[S(n)]$，即：$\theta>0$，称 θ 为安全系数；

（3）$S(n)=W_1+W_2+\cdots+W_n$ 表示前面 n 期的理赔总量，$W_i(i=1,2,\cdots,n)$ 为第 i 年的总理赔量，一般情况下 $W_1,W_2,\cdots,W_n$ 是独立同分布的随机变量，共同分布记为 $F_W(x)$。

同样地，对于离散时间盈余过程（7.5.1）也会考虑与破产有关的各种变量。

定义 7－13 对盈余序列（7.5.1）定义如下与破产有关的变量：

称 $\tilde{T}=\min\{n:U(n)<0\}$ 为破产时刻；

称 $\tilde{\phi}(u,n)=P(U(t)\geqslant0,\forall t=0,1,\cdots,n),n=1,2,\cdots$，为离散时间有限生存概率；

称 $\tilde{\psi}(u,n)=1-\tilde{\phi}(u,n)=P(\exists t=1,2,\cdots,n,U(t)<0),n=1,2,\cdots$，为离散时间有限破产概率；

称 $\tilde{\phi}(u)=P(U(t)\geqslant0,\forall t=0,1,\cdots),n=1,2,\cdots$，为离散时间终极生存概率；

称 $\tilde{\psi}(u)=1-\tilde{\phi}(u)=P(\exists t=0,1,\cdots,U(t)<0),n=1,2,\cdots$，为离散时间终极破产概率；

性质 7－2 对于 $u_1\leqslant u_2$ 和 $0<n_1\leqslant n_2<\infty$，有以下结论成立：

（1）$\tilde{\psi}(u_2,n)\leqslant\tilde{\psi}(u_1,n),n=1,2,\cdots$；

（2）$\tilde{\psi}(u_2)\leqslant\tilde{\psi}(u_1)$；

（3）$\tilde{\psi}(u,n_1)\leqslant\tilde{\psi}(u,n_2)\leqslant\tilde{\psi}(u)$；

（4）$\lim\limits_{n\to\infty}\tilde{\psi}(u,n)=\tilde{\psi}(u)$。

离散时间盈余序列破产概率的计算不易像连续时间模型的定理 7－1 那样考虑微积分方程。但是，可以得到一些关于时间的递推关系。

性质 7－3 盈余序列（7.5.1）的离散时间有限生存概率 $\tilde{\phi}(u,n)$ 满足以下递推关系：

(1) $\tilde{\phi}(u,n)=\tilde{\phi}(u,n-1)\times P(U(n)\geqslant 0\mid U(n-1)\geqslant 0)$,

$n=1,2,\cdots$, $\tilde{\phi}(u,0)=1$

(2) $\tilde{\psi}(u,n)=P(U(n)<0\mid U(n-1)\geqslant 0)+\tilde{\psi}(u,n-1)\times P(U(n)\geqslant 0\mid U(n-1)\geqslant 0)$

只要利用定义 7－8 中的表达式

$$\tilde{\phi}(u,n)=P\{U(1)\geqslant 0,\quad U(2)\geqslant 0,\cdots,U(n)\geqslant 0\},\quad n=1,2,\cdots$$

和盈余过程的马氏性很容易证明性质 7－3。基于这个性质，可以得到生存概率关于时间的积分递推公式。

定理 7－9　将盈余序列（7.5.1）的生存概率 $\tilde{\phi}(u,n)$ 看做 (u,n) 的二元函数，则有限生存概率满足以下递推公式：

$$\tilde{\phi}(u,n)=\int_0^{u+c}\tilde{\phi}(u+c-x,n-1)dF_W(x),n=1,2,\cdots,u\geqslant 0 \tag{7.5.2}$$

特别地，有：$\tilde{\phi}(u,1)=F_W(u+c)$。

【例 7－10】　现有如下盈余序列：初始盈余为 2，年保费收入为 3，年损失量分布如下：$P(W=0)=0.6=1-P(W=6)$。试计算前两年内破产的概率 $\tilde{\psi}(2,2)$。

解：

$$U(0)=2,U(1)=\begin{cases}5, & p=0.6\\ -1, & q=0.4\end{cases}$$

所以：　$\tilde{\phi}(2,1)=0.6$

又有　$P(U(1)=5,U(2)=8)=0.6\times 0.6$

$P(U(1)=5,U(2)=2)=0.6\times 0.4$

$P(U(1)=-1,U(2)=2)=0.4\times 0.6$

$P(U(1)=-1,U(2)=-4)=0.4\times 0.4$

所以，$\tilde{\phi}(2,2)=P(U(1)>0,U(2)>0)=0.6\times 0.6+0.6\times 0.4=0.6$。因此，$\tilde{\psi}(2,2)=0.4$。但是，注意：

$$P(U(2)>0)=0.6\times 0.6+0.6\times 0.4+0.4\times 0.6=0.84\neq\tilde{\phi}(2,2)$$

■

类似地，可以定义盈余序列模型的调节系数 $\tilde{R}$ 为方程

$$e^{-cr}M_W(r)=1 \tag{7.5.3}$$

的正根。注意，式（7.5.3）可以变形为：

$$\log M_W(r) - cr = 0 \tag{7.5.4}$$

当 W_i 为复合泊松分布时，有 $\log M_W(r) = \lambda[M_X(r) - 1]$，所以，式（7.5.4）与 $\lambda M_X(r) = \lambda + cR$ 等价，即当第 i 个时期的总理赔量为复合泊松分布时，离散时间模型的调节系数与连续时间的调节系数相同。

基于调节系数方程（7.5.3），也可以得到一个与定理 7－8 相类似的结论。

定理 7－10 对于定义 7－7 给出的盈余序列，当 $u \geqslant 0$ 时，破产概率 $\tilde{\psi}(u)$ 有以下的表达式：

$$\tilde{\psi}(u) = \frac{e^{-\tilde{R}u}}{E[e^{-\tilde{R}U(\tilde{T})} \mid \tilde{T} \leqslant \infty]}, \qquad u \geqslant 0 \tag{7.5.5}$$

其证明过程可参阅（Bower etc，1997）。■

由定理 7－10 自然有

$$\tilde{\psi}(u) < e^{-\tilde{R}u}, \qquad u \geqslant 0$$

成立。

【例 7－11】 若 W_i 服从正态分布 $N(\mu, \sigma^2)$，求调节系数 $\tilde{R}$ 的表达式。

解： 由已知：

$$\log M_W(r) = \mu r + \frac{1}{2}\sigma^2 r^2$$

将其代入方程（7.5.4），有

$$\tilde{R} = \frac{2(c-\mu)}{\sigma^2}$$

另外，对于一般的 $F_W(x)$，有 $\log M_W(r)$ 的展开式：

$$\log M_W(r) = \mu r + \frac{1}{2}\sigma^2 r^2 + \cdots$$

其中，$\mu = E[W]$，$\sigma^2 = Var[W]$。因此 $\tilde{R} \cong \dfrac{2(c-\mu)}{\sigma^2}$ 也可看做是对 $\tilde{R}$ 的一种近似。■

特别地，当 W_i 为随机变量 N 与 X 的复合分布，且 $c = (1+\theta)\mu$，$\tilde{R}$ 可近似地表示为：

$$\tilde{R} \approx \frac{2\theta p_1 E(N)}{(p_2 - p_1^2)E(N) + p_1^2 Var(N)} \tag{7.5.6}$$

其证明过程比较简单，读者可自行证明。

【例 7－12】 在以下情形计算 $\tilde{R}$ 的近似值：

（1）N 服从参数为 λ 的泊松分布，$\lambda > 0$；

（2）N 服从参数为 r 和 p 的负二项分布，$r>1$，$0<p<1$。

解：

（1）将 $E(N)=Var(N)=\lambda$ 代入（7.5.5），有

$$\tilde{R}\approx\frac{2\theta p_1}{p_2}$$

（2）将 $E(N)=\frac{rq}{p}$和 $Var(N)=\frac{rq}{p^2}$ 代入式（7.5.5），有

$$\tilde{R}=\frac{2\theta p_1}{p_2+p_1^2\left(\frac{1}{p}-1\right)}$$

当 $p\to1$ 时，将得到与（1）相同的结果。 ■

§7.6　最优再保险与调节系数

我们在第六章已经了解，按照原保险人与再保险人分担损失的方式，可以把再保险分为限额损失再保险和比例再保险。对于比例再保险，由于其损失分割和保费摊回都按照固定比例确定，因此再保险的风险与收益比较明晰，相对容易决策；而限额损失再保险却由于自留额与承保的损失密切相关，因此必须在综合评估风险的基础上才能确定自留额。

这一节将通过两个例子说明，首先，在其他条件相同的情况下，限额损失再保险的风险小于比例再保险；其次，如何确定限额损失再保险的自留额。与第六章从平均损失的角度讨论风险不同，这一节我们将从调节系数的角度来评估风险。

7.6.1　再保险附加费率与调节系数

我们对限额损失再保险模型和比例损失再保险模型从破产的角度作一个比较分析，这里假定所有的再保险规则都是针对原保险人的理赔额变量 X 设定的，则再保险人承担的损失如下：

$$\text{限额损失再保险：}I_d(X)=\begin{cases}0, & X\leqslant d\\ X-d, & X>d\end{cases}\tag{7.6.1}$$

比例再保险：$I(X)=kX$，$0<k<1$，k 为原保险人的分出比例

（7.6.2）

【例 7－13】 已知理赔总量为复合泊松过程 $S(t)$，参数 $\lambda=1$，理赔额 X 在［0，1］上均匀分布，保险人的费率为 $c=1$。现在考虑再保险附加费率为 100% 和 140% 两种情况：

（1）比例再保险模型：$I(X)=kX$，分别对 $k=0$，0.1，0.2，…，1 计

算调节系数；

（2）限额再保险模型：$I(X)=I_d(X)$，分别对 $d=0$，0.1，0.2，…，1 计算调节系数；

（3）在期望收益相等的条件下，比较两类模型的调节系数与再保险安排的关系。

解：

因为 $p_1=\frac{1}{2}$，$\lambda=1$，而且 $c=1$，所以原保险人的附加费率为 $\theta=1$。

① 对于不同的 k，设相应的再保险费率为 c_k，θ 为再保险附加费率，则有

$$c_k=(1+\theta)\lambda E[I(X)]=(1+\theta)\lambda kE[X]=\frac{(1+\theta)k}{2}$$

原保险人的自留保费为：$1-\frac{(1+\theta)k}{2}$，实施再保后原保险人自留风险的期望为：

$$\lambda(1-k)E[X]=\frac{1-k}{2}$$

再保后的调节系数方程为：$\lambda+(c-c_k)r=\lambda M_X((1-k)r)$，代入后有：

$$1+\frac{2-k-\theta k}{2}r=\frac{e^{r(1-k)}-1}{r(1-k)}$$

当 $\theta=1$ 时，再保后的调节系数方程为：

$$1+(1-k)r=\frac{e^{r(1-k)}-1}{r(1-k)} \tag{7.6.3}$$

若记 $\tilde{r}=(1-k)r$，且 $\tilde{R}$ 为

$$e^{\tilde{r}}=1+\tilde{r}+\tilde{r}^2 \tag{7.6.4}$$

的非零解，式（7.6.4）相当于 $k=0$ 时（不安排再保）的调节系数方程。

则对于一般的 $k=0$，0.1，0.2，…，1，式（7.6.3）的解为：$R=\frac{\tilde{R}}{1-k}$。

经过数值计算，（7.6.4）的解为 1.793，所以，$\theta=1$ 时，式（7.6.3）中再保后的调节系数为 $R=1.793/(1-k)$，可以看出，R 是比例再保险中再保比例系数 k 的增函数，也就是说，随着再保比例的增加，原保险人的调节系数也会增加，破产概率的上界会降低，这与现实的直觉相符。这个例题中，出现这种再保比例系数与调节系数直接的单调增关系，可能是由于原保险的附加费率与再保的附加费率相同，均为 1。

对应不同的 $k=0$，0.1，0.2，…，1 值，式（7.6.3）的解列在表 7-4 中。

表 7-4　　例 7-13 的调节系数（1）

k	调节系数 R		d	调节系数 R	
	$\theta=1$	$\theta=1.4$		$\theta=1$	$\theta=1.4$
0.0	1.793	1.793	1.0	1.793	1.793
0.1	1.993	1.936	0.9	1.833	1.828
0.2	2.242	2.095	0.8	1.940	1.920
0.3	2.562	2.268	0.7	2.116	2.062
0.4	2.989	2.436	0.6	2.378	2.259
0.5	3.587	2.538	0.5	2.768	2.518
0.6	4.483	2.335	0.4	3.373	2.840
0.7	5.978	0.635	0.3	4.400	3.138
0.8	8.966	…	0.2	6.478	2.525
0.9	17.933	…	0.1	12.746	…
1.0	∞	…	0.0	∞	…

当 $\theta=1.4$ 时，原保险人的自留保费为：$1-\frac{(1+\theta)k}{2}=1-1.2k$

实施再保后原保险人自留风险的期望为：$\lambda(1-k)E[X]=(1-k)/2$，再保后的调节系数方程为：

$$1+(1-1.2k)r=\frac{e^{r(1-k)}-1}{r(1-k)} \tag{7.6.5}$$

式（7.6.5）没有与式（7.6.3）相似的一般性结果，只能分别对 $k=0$，0.1，0.2，…，1 计算数值解，结果列入表 7-3。从计算结果可以发现，调节系数 R 先是随着再保险比例 k 的增加而缓慢上升，也就是说破产概率降低；但是，当 k 上升到一定水平（$k>0.5$）后，调节系数 R 急剧下降，也就是说破产概率增加，原因是随着 k 的增加原保险人的自留保费 $1-1.2k$ 可能不足以支付原保险人自留风险的期望（$1-k$）/2，即原保险人（从最终的意义看）必然破产，调节系数为零或者是负值。

特别地，当原保险人的自留保费 $1-1.2k$ 与原保险人自留风险的期望 $(1-k)/2$ 相等时，有：$k=5/7$。调节系数方程（7.6.5）为：

$$e^{r(1-k)}=1+r(1-k)+\frac{1}{2}[r(1-k)]^2$$

由函数 e^x 的性质知，上式在非负的定义域内没有非零解。所以，当 $k=0.8$，0.9，1 时，不存在非负的调节系数。

② 在限额损失再保险模型下，设相应的再保险费率为 c_d，则

$$c_d=(1+\theta)\lambda E(I_d)=(1+\theta)\lambda\int_d^1(x-d)dx=\frac{(1+\theta)(1-d)^2}{2}$$

原保险人的自留保费为：$1-\frac{(1+\theta)(1-d)^2}{2}$，实施再保后原保险人自留风险的期望为：$\lambda(E[X]-E[I_d])=\frac{1-(1-d)^2}{2}$。

分别对两种再保附加费率的情况确定调节系数方程：

$$\theta=1\text{ 时}，1+[1-(1-d)^2]r=\frac{e^{rd}-1}{r}+(1-d)e^{rd}$$

$$\theta=1.4\text{ 时}，1+[1-1.2(1-d)^2]r=\frac{e^{rd}-1}{r}+(1-d)e^{rd}$$

这两个表达式没有一般的解，只能数值求解。结果列在表 7－3 中。

与比例再保的情形类似，当 $\theta=1$ 时，随着自留额 d 的降低，原保险人的调节系数增加，破产概率的上界降低。当 $\theta=1.4$ 时，从计算结果，我们发现，调节系数 R 先是随着自留额 d 的降低而缓慢上升，也就是说破产概率降低；但是，当 d 降到一定水平（$k<0.3$）后，调节系数 R 急剧下降，也就是说破产概率增加，原因是随着 d 的降低原保险人的自留保费 $1-\frac{(1+\theta)(1-d)^2}{2}=1-1.2(1-d)^2$ 可能不足以支付原保险人的自留风险的期望 $\frac{1-(1-d)^2}{2}$，也就是说原保险人（从最终的意义看）必然破产，调节系数为零或者是负值。

特别地，当原保险人的自留保费与原保险人自留风险的期望相等时，有：$d=1-\sqrt{5/7}=0.1548$，调节系数方程在非负的定义域内没有非零解。所以，当 $d=0.1$ 和 $d=0$ 时，不存在非负的调节系数。

③ 当附加费率相同时，要保持不同再保方式下有相同的期望收益，只要令：

$$E(kX)=E[I_d(X)]$$

即 k 与 d 满足 $k=(1-d)^2$，然后固定 $d=0$，0.1，0.2，…，1，由上式解得对应的 k 值，并进一步求解调节系数，最终结果列入表 7－5。

表 7－5　　例 7－13 的调节系数（2）

k	d	调节系数 R			
		$\theta=1$		$\theta=1.4$	
		比例损失模型	限额损失模型	比例损失模型	限额损失模型
0.00	1.0	1.793	1.793	1.793	1.793
0.01	0.9	1.811	1.833	1.807	1.828
0.04	0.8	1.868	1.940	1.848	1.920
0.09	0.7	1.971	2.116	1.921	2.062
0.16	0.6	2.135	2.378	2.030	2.259
0.25	0.5	2.391	2.768	2.181	2.518
0.36	0.4	2.802	3.373	2.372	2.840
0.49	0.3	3.516	4.400	2.535	3.138
0.64	0.2	4.981	6.478	1.992	2.525
0.81	0.1	9.438	12.746	…	…
1.00	0.0	∞	∞	…	…

计算结果表明，在两类再保险模型期望收益相等的条件下时，无论再保险的附加费率如何选取，限额损失再保险的调节系数始终大于比例再保险的调节系数，在一定的意义下，也就是说，限额损失再保险破产的可能性相对较低，风险较小。实际上，这一结果更一般的结论与效用函数有关，此处不再赘述。 ■

7.6.2 自留额与调节系数

通过第六章的例 6－11 我们知道，随着原保险人自留额的提高，再保险人的承保风险逐渐降低。下面我们从调节系数的角度再次考虑确定自留额的问题。

【例 7－14】 已知某保单组合每年发生的聚合理赔量是相互独立且具有相同分布的随机变量，均为复合泊松分布，参数为 $\lambda=1.5$，理赔额变量 X 服从分布 $p(1)=2/3=1-p(2)$，又假设各年的保费均按照费率 $c=2.5$ 收取。保险人现在考虑选择限额损失再保险方式对每年的总理赔进行再保险，如果再保险人按 100% 的附加费率收取再保费，试计算下列情况下原保险人自留风险的调节系数及期望收益：

（1） 不进行再保；

（2） 免赔额为 0，1，2，…，10。

解： 这时，各年的总理赔模型与例 6－11 相同。对于 $d=3,4,5,\cdots$，记 W_d 为原保险人各年自留风险的总理赔额变量，W_d 的分布为：

$$P(W_d=x)=f_S(x),x=0,1,\cdots,d-1\ ;P(W_d=d)=1-F_S(d-1)$$

其中 $f_S(x)$ 和 $F_S(x)$ 均来自表 6－7。

原保险人各年的自留保费记为 c_d，则 $c_d=c-2E(I_d)=2.5-2E(I_d)$。调节系数方程为：

$$e^{c_d r}=M_{W_d}(r)=\sum_{x=0}^{d-1}f_S(x)e^{xr}+[1-F_S(d-1)]e^{dr}$$

保险人自留风险的期望收益为：

$$c_d-E(W_d)=2.5-2E(I_d)-[E(W)-E(I_d)]=0.5-E(I_d)$$

所以，由表 6－7 可知，当 $d\leqslant 2$ 时，不会实施限额损失再保险。

对于不进行再保险的情况，由式（7.4.1）得：

$$1.5+2.5r=e^r+\frac{1}{2}e^{2r}$$

经过数值计算，解得 $R=0.28$，此时保险人的期望收益为 0.5。

根据表 6－7 的数据和上述的计算过程，我们将各种免赔额的计算结果列入表 7－6（请读者自行验证，特别注意调节系数的计算）。

表 7－6　　　　例 7－14 的计算结果

免赔额 d	自留保费 c_d	调节系数	期望收益
3	1.822	0.249	0.161
4	2.186	0.347	0.343
5	2.364	0.336	0.432
6	2.445	0.315	0.473
7	2.479	0.301	0.489
8	2.492	0.292	0.496
9	2.497	0.287	0.499
10	2.499	0.284	0.500

从以上的计算结果可以看出，随着自留额的增加期望收益逐渐增加，所以不分保的期望收益最大。如果我们简单地将调节系数等同于破产概率，则调节系数越大破产概率的上界越小，也可以看做是破产概率越小（并不总是成立）。从上面的计算中可以发现，调节系数并不是随着自留额的增加而愈来愈大，也就是说，并不是像直观感觉的那样：自留的风险越多破产概率越大。当免赔额为 4 时，调节系数最大，也可以看做是破产概率最小，或者说破产风险最小、最安全；而免赔额为 3 时调节系数很小、期望收益也很小，是最不可取的再保方式，免赔额为 5 时，调节系数不如免赔额为 4 的大，但是期望收益有所提高。因此，从破产风险与收益的平衡看，免赔额为 4 或 5 都是可选的再保方式。 ■

§7.7　布朗运动与盈余过程

在这一节之前，我们对破产概率所得到的最好的结果就是式（7.4.7），但是这个结果比较复杂而且不利于计算。在一定的条件下，我们可以把净现金流入过程近似为一个带漂移项的布朗运动，从而得到破产概率更为简洁的表达。

7.7.1　布朗运动与净现金流入过程

定义 7－14　称 $\{W(t),\ t\geqslant 0\}$ 为带漂移的布朗运动，如果

（1）$W(0)=0$；

（2）$\{W(t),\ t\geqslant 0\}$ 具有平稳独立增量性；

（3）对任意的 $t>0$，有 $W(t)\sim N(\mu t,\sigma^2 t)$，其中 $\mu\geqslant 0$。

特别地，当 $\mu=0$ 时称 $\{W(t),\ t\geqslant 0\}$ 为布朗运动，布朗运动的每条轨道处处连续，处处不可导。

我们重新考虑 7.2 节讨论的盈余过程：

$$U(t)=u+ct-S(t)=u+ct-\sum_{i=1}^{N(t)}X_i,t\geqslant 0$$

其中 u 为初始盈余，$u \geqslant 0$；c 为单位时间内收取的保费；$\{N(t), t \geqslant 0\}$ 是强度为 λ 的泊松过程，理赔额变量 X_i，$i=1, 2, \cdots$ 独立同分布，都与 N_t 独立，X_i 都取正值，且矩母函数存在。

盈余过程是一个跳跃变化的过程，它从 $U(0)=u$ 出发，以速率 c 连续递增，并且在索赔发生时间 $\{T_1, T_2, \cdots\}$ 发生向下的跳跃 $X_1, X_2, \cdots$，盈余过程的一个轨道见图 7－5，其中 $U(t)=30+35t-\sum_{i=1}^{N(t)} X_i$，$t \geqslant 0$，$\lambda=4$，$X$ 服从均值为 8 的指数分布。

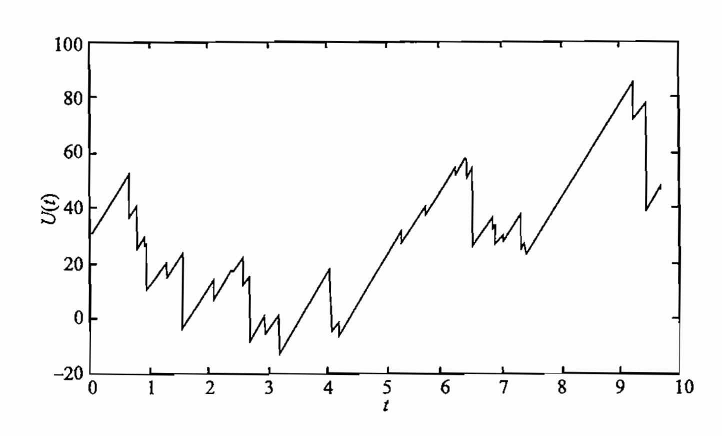


图 7－5 泊松盈余过程的样本轨道

如果令 $Z(t)=ct-S(t)$，$t \geqslant 0$，则

$$U(t)=u+Z(t), \quad t \geqslant 0$$

$Z(t)$ 是时间段 $[0, t]$ 内保险公司的净现金流入。对于承保范围很广的险种，投保与索赔通常很频繁，而且相对都是小额索赔，也就是说，c 与 λ 相当大，而 X 却比较小，此时，过程 $\{Z(t), t \geqslant 0\}$ 的轨道频繁地进行小幅的跳跃，如图 7－6 所示，其中 $c=35$，$\lambda=1\,000$，X 服从均值为 0.03 的指数分布，这与带漂移项的布朗运动具有极小均值和极小方差的特征非常相似，这启发我们利用带漂移项的布朗运动来近似净现金流入过程。

定理 7－11 当 $\lambda \to \infty$ 时，满足上述条件的净现金流入过程 $\{Z(t), t \geqslant 0\}$ 可以表示为带漂移项的布朗运动。

证明： 首先有 $Z(0)=0$，而且由 $S(t)$ 的平稳、独立增量的特征易知 $Z(t)$ 也满足增量的平稳性和独立性。

下面考虑其近似分布，由于随机变量的矩母函数具有与分布一一对应的特征，我们从考察 $Z(t)$ 的矩母函数 $M_{Z(t)}(r)$ 入手。

因为
$$M_{Z(t)}(r)=M_{ct-S(t)}(r)=E[e^{r(ct-S(t))}]$$
$$=e^{trc}E[e^{-S(t)r}]=e^{tcr}M_{S(t)}(-r)=e^{tcr}e^{\lambda t[M_X(-r)-1]}$$

所以，
$$\frac{\ln M_{Z(t)}(r)}{t}=cr+\lambda[M_X(-r)-1] \tag{7.7.1}$$

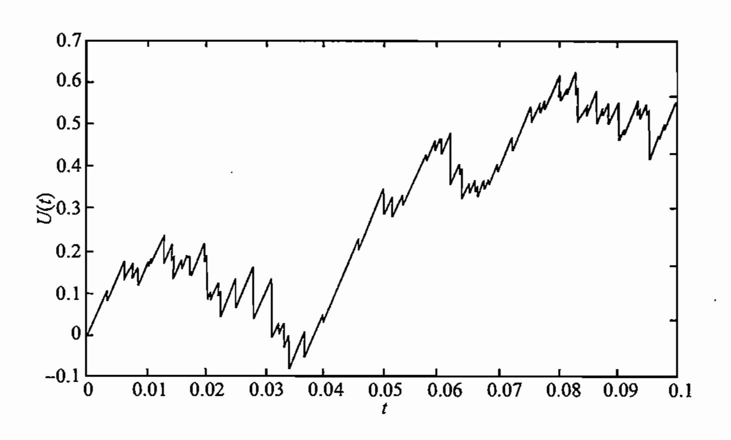


图 7-6　净现金流入过程的样本轨道

由已知 $E[Z(t)]=[c-\lambda E(X)]t, Var[Z(t)]=\lambda E(X^2)t$

令 $c-\lambda EX=\mu$，$\lambda E(X^2)=\sigma^2$ 那么，有 $c=\mu+\lambda EX$，将其代入式 (7.7.1)，有

$$\frac{\ln M_{Z(t)}(r)}{t}=[\mu+\lambda E(X)]r+\lambda[M_X(-r)-1] \tag{7.7.2}$$

再把 $M_X(-r)$ 在 0 点进行泰勒展开并代入式 (7.7.2)，有

$$\begin{aligned}\frac{\ln M_{Z(t)}(r)}{t}&=[\mu+\lambda E(X)]r+\lambda[1-rE(X)+\frac{r^2}{2!}E(X^2)\\&\quad-\frac{r^3}{3!}E(X^3)+\cdots-1]\\&=\mu r+\frac{r^2}{2}\lambda E(X^2)-\lambda[\frac{r^3}{3!}E(X^3)-\frac{r^4}{4!}E(X^4)+\cdots]\\&=\mu r+\frac{\sigma^2}{2}r^2-\lambda[\frac{r^3}{3!}E(X^3)-\frac{r^4}{4!}E(X^4)+\cdots]\end{aligned} \tag{7.7.3}$$

再令理赔额随机变量 $X=\alpha Y$，其中 Y 具有固定的均值和方差，那么，

$$E(X^k)=\alpha^k E(Y^k),\qquad k=0,1,2,\cdots$$

所以式 (7.7.3) 变为：

$$\begin{aligned}\frac{\ln M_{Z(t)}(r)}{t}&=\mu r+\frac{\sigma^2}{2}r^2-\lambda\alpha^2\left[\frac{\alpha r^3}{3!}E(Y^3)-\frac{\alpha^2 r^4}{4!}E(Y^4)+\cdots\right]\\&=\mu r+\frac{\sigma^2}{2}r^2-\sigma^2\left[\frac{\alpha r^3 E(Y^3)}{3!\ E(Y^2)}-\frac{\alpha^2 r^4 E(Y^4)}{4!\ E(Y^2)}+\cdots\right]\end{aligned} \tag{7.7.4}$$

再由 $\lambda=\frac{\sigma^2}{E(X^2)}=\frac{\sigma^2}{E(Y^2)}\cdot\frac{1}{\alpha^2}$ 可知 $\alpha\to0$ 时，$\lambda\to\infty$，

因此，$\lim\limits_{\alpha\to0}\left[\frac{\alpha r^3 E(Y^3)}{3!\ E(Y^2)}-\frac{\alpha^2 r^4 E(Y^4)}{4!\ E(Y^2)}+\cdots\right]=0$

等价于 $\lim\limits_{\lambda\to\infty}\left[\frac{\alpha r^3 E(Y^3)}{3!\ E(Y^2)}-\frac{\alpha^2 r^4 E(Y^4)}{4!\ E(Y^2)}+\cdots\right]=0$

所以有　$$\lim_{\lambda\to\infty}\frac{\ln M_{Z(t)}(r)}{t}=\mu r+\frac{\sigma^2}{2}r^2 \tag{7.7.5}$$

即　$$\lim_{\lambda\to\infty}M_{Z(t)}(r)=e^{\mu tr+\frac{\sigma^2 t}{2}r^2} \tag{7.7.6}$$

式（7.7.6）的右端就是均值为 μt、方差为 $\sigma^2 t$ 的正态分布的矩母函数，这就证明了净现金流入过程的极限过程是带漂移项的布朗运动。 ■

7.7.2　盈余过程的破产概率

既然净现金流入过程在一定条件下可以用带漂移项的布朗运动近似，那么我们就可以假定初始资本为 u 的盈余过程为：

$$U(t)=u+W(t),t\geqslant 0 \tag{7.7.7}$$

其中 $\{W(t),\ t\geqslant 0\}$ 为带漂移的布朗运动，其均值和方差分别为 μt 和 $\sigma^2 t$。

下面我们再考虑有限时间区间（0，τ）内的破产概率 $\psi(u,\ \tau)$：

$$\psi(u,\tau)=P(T<\tau)=P(\min_{0<t<\tau}U(t)<0)=P(\min_{0<t<\tau}W(t)<-u)$$

定理 7－12　对于上述描述的盈余过程，有以下结论成立：

（1）有限时间破产概率为：

$$\psi(u,\tau)=\Phi\left(-\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right)+e^{-\frac{2\mu}{\sigma^2}u}\cdot\Phi\left(-\frac{u-\mu\tau}{\sqrt{\sigma^2\tau}}\right) \tag{7.7.8}$$

其中 Φ（·）是标准正态分布的累积分布函数；

（2）终极破产概率为：

$$\psi(u)=e^{-\frac{2\mu}{\sigma^2}u} \tag{7.7.9}$$

（3）破产时刻 T 的条件分布函数为：

$$F_{T|T<\infty}(\tau)=e^{\frac{2\mu}{\sigma^2}u}\cdot\Phi\left(-\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right)+\Phi\left(-\frac{u-\mu\tau}{\sqrt{\sigma^2\tau}}\right),\tau>0 \tag{7.7.10}$$

证明：

（1）对于盈余过程 $\{U(t),\ t\geqslant 0\}$ 在时间区间（0，τ）内的任意一条样本轨道，它在 $t=\tau$ 的值无非是 $U(\tau)<0$ 和 $U(\tau)\geqslant 0$ 两种情况。如果 $U(\tau)<0$，那么相应的样本轨道必在某一时刻 $T<\tau$ 处穿过时间轴，从而存在使 $U(t)<0$ 的点，即破产发生。我们令这类样本轨道的集合为：

$$A_n^-=\{T\leqslant\tau,-n\delta<U(\tau)\leqslant-(n-1)\delta\},\qquad \delta>0,n=1,2,\cdots \tag{7.7.11}$$

由于 $U(\tau)<0$ 时必有破产发生，因此，

$$A_n^-=\{-n\delta\leqslant U(\tau)<-(n-1)\delta\}$$

对于 $U(\tau)\geqslant 0$ 的样本轨道，我们需要考虑那些破产发生的情况，令

$$A_n^+=\{T\leqslant\tau,(n-1)\delta\leqslant U(\tau)<n\delta\},\qquad \delta>0,n=1,2,\cdots \tag{7.7.12}$$

它是那些破产发生并且 $U(\tau)\geqslant 0$ 的样本轨道的集合。

因此，

$$P(T \leqslant \tau) = P(T \leqslant \tau, U(\tau) < 0) + P(T \leqslant \tau, U(\tau) > 0) \tag{7.7.13}$$

$$= P(U(\tau) < 0) + \sum_{n=1}^{\infty} P(A_n^+) \tag{7.7.14}$$

$$= P(U(\tau) < 0) + \sum_{n=1}^{\infty} P(A_n^-) \cdot \frac{P(A_n^+)}{P(A_n^-)}$$

而由已知 $U(\tau) \sim N(u+\mu\tau, \sigma^2\tau)$，所以，

$$P(U(\tau) < 0) = P\left(\frac{U(\tau) - (u+\mu\tau)}{\sqrt{\sigma^2\tau}} < -\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right) = \Phi\left(-\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right) \tag{7.7.15}$$

又因为

$$P(A_n^+) = \int_0^{\tau} P((n-1)\delta \leqslant U(\tau) < n\delta \mid T = s)\, dF_T(s)$$

$$= \int_0^{\tau} P((n-1)\delta \leqslant U(\tau) - U(s) < n\delta)\, dF_T(s)$$

$$= \int_0^{\tau} \int_{(n-1)\delta}^{n\delta} \frac{1}{\sqrt{2\pi\sigma^2(\tau - s)}} e^{-\frac{[x-\mu(\tau-s)]^2}{2\sigma^2(\tau-s)}} dx dF_T(s)$$

$$= e^{\frac{n\delta\mu}{\sigma^2}} \int_0^{\tau} \frac{1}{\sqrt{2\pi\sigma^2(\tau - s)}} e^{-\frac{(n\delta)^2+[\mu(\tau-s)]^2}{2\sigma^2(\tau-s)}} dF_T(s) + o(\delta) \tag{7.7.16}$$

同理，

$$P(A_n^-) = e^{-\frac{n\delta\mu}{\sigma^2}} \int_0^{\tau} \frac{1}{\sqrt{2\pi\sigma^2(\tau - s)}} e^{-\frac{(n\delta)^2+[\mu(\tau-s)]^2}{2\sigma^2(\tau-s)}} dF_T(s) + o(\delta) \tag{7.7.17}$$

所以，

$$\frac{P(A_n^+)}{P(A_n^-)} = e^{\frac{2n\delta\mu}{\sigma^2}} + o(\delta) \tag{7.7.18}$$

于是，

$$\sum_{n=1}^{\infty} P(A_n^+) = \sum_{n=1}^{\infty} P(A_n^-)[e^{\frac{2n\delta\mu}{\sigma^2}} + o(\delta)]$$

$$= \sum_{n=1}^{\infty} [\int_{-n\delta}^{-(n-1)\delta} \frac{1}{\sqrt{2\pi\sigma^2\tau}} e^{-\frac{(x-\mu\tau-u)^2}{2\sigma^2\tau}} dx] \cdot [e^{\frac{2n\delta\mu}{\sigma^2}} + o(\delta)]$$

$$= \sum_{n=1}^{\infty} [\int_{-n\delta}^{-(n-1)\delta} \frac{1}{\sqrt{2\pi\sigma^2\tau}} e^{-\frac{(x-\mu\tau-u)^2}{2\sigma^2\tau} - \frac{2\mu x}{\sigma^2}} dx] \cdot [1 + o(\delta)] \tag{7.7.19}$$

在式（7.7.19）中令 $\delta \to 0$，得

$$\sum_{n=1}^{\infty} P(A_n^+) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\sigma^2\tau}} e^{-\frac{(x-\mu\tau-u)^2+4\mu x\tau}{2\sigma^2\tau}} dx$$

$$= e^{-\frac{2\mu}{\sigma^2}u} \cdot \Phi\left(-\frac{u-\mu\tau}{\sqrt{\sigma^2\tau}}\right) \tag{7.7.20}$$

把式（7.7.15）和式（7.7.20）代入式（7.7.14），可得

$$\psi(u,\tau)=\Phi\left(-\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right)+e^{-\frac{2\mu}{\sigma^2}u}\cdot\Phi\left(-\frac{u-\mu\tau}{\sqrt{\sigma^2\tau}}\right)$$

（2）在式（7.7.8）中令 $\tau\to\infty$ 即可得终极破产概率为 $\psi(u)=e^{-\frac{2\mu}{\sigma^2}u}$。

（3）破产时刻的条件分布函数为

$$F_{T|T<\infty}(\tau)=P(T<\tau|T<\infty)=\frac{\psi(u,\tau)}{\psi(u)}$$

$$=e^{\frac{2\mu}{\sigma^2}u}\cdot\Phi\left(-\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right)+\Phi\left(-\frac{u-\mu\tau}{\sqrt{\sigma^2\tau}}\right),\tau>0$$

其概率密度函数为：

$$f_{T|T<\infty}(\tau)=\frac{u}{\sqrt{2\pi\sigma^2}}\pi^{-\frac{3}{2}}\cdot e^{-\frac{(u-\mu\tau)}{2\sigma^2\tau}},\tau>0 \qquad (7.7.21)$$

■

【例 7-15】 某泊松盈余过程的初始盈余为 u，$u\geqslant0$，保费附加因子为 θ，泊松强度参数为 λ，理赔额变量 X 存在有限的均值和方差，试在布朗运动近似的条件下写出其有限时间破产概率、终极破产概率以及破产时刻的概率密度函数。

解： 设此泊松盈余过程单位时间内所收的保费为 c，其净现金流入过程为 $\{Z(t),\ t\geqslant0\}$，由已知，有

$$E[Z(t)]=[c-\lambda E(X)]t=\theta\lambda E(X)t$$

$$Var[Z(t)]=\lambda E(X^2)t$$

令 $\mu=\theta\lambda E(X)$，$\sigma^2=\lambda E(X^2)$ 并将其代入式（7.7.8）、（7.7.9）和（7.7.10）中，则有限时间破产概率为：

$$\psi(u,\tau)=\Phi\left(-\frac{u+\theta\lambda E(X)\tau}{\sqrt{\lambda E(X^2)\tau}}\right)+e^{-\frac{2E(X)}{E(X^2)}\theta u}\cdot\Phi\left(-\frac{u-\theta\lambda E(X)\tau}{\sqrt{\lambda E(X^2)\tau}}\right),$$

$$u>0,\tau>0$$

终极破产概率为：

$$\psi(u)=e^{-\frac{2E(X)}{E(X^2)}\theta u},\qquad u>0,\tau>0$$

破产时刻 T 的条件概率密度函数为：

$$f_{T|T<\infty}(\tau)=\frac{u}{\sqrt{2\pi\lambda E(X^2)}}\pi^{-\frac{3}{2}}\cdot e^{-\frac{(u-\theta\lambda E(X)\tau)}{2\lambda E(X^2)\tau}},\qquad \tau>0$$

■

【例 7-16】 已知泊松盈余过程的平均破产时刻为 10，理赔额变量的密度函数为 $f(x)=(1+6x)e^{-3x}$，$x\geqslant0$，保费附加因子为 18%，泊松强度参数为 5 980，求初始盈余的近似值。

解： 由附录可知式（7.7.21）是均值为 u/μ，方差为 $u\sigma^2/\mu^3$ 的逆高斯分布的概率密度函数，即 $ET=u/\mu=10$，再由已知

$$\theta = 0.18, \lambda = 5\,980, E(X) = \int_0^\infty x(1+6x)e^{-3x}dx = \frac{5}{9}$$

所以 $\mu = \theta\lambda E(X) = 598$，所以 $u = 5\,980$。

习　题

1. 试证明性质 7－1 和性质 7－2。

2. 设理赔过程为复合泊松过程，泊松参数为 λ，理赔额变量为伽玛分布 Gamma(α, β)，试求理赔过程的矩母函数。

3. 对于定理 7－2 中的式（7.3.5），验证破产概率 $\psi(u)$ 是附加费率 θ 和初始资金 u 的递减函数。

4. 设盈余过程中的理赔过程为复合泊松过程，理赔额变量服从均值为 1 的指数分布，$c=4$；又设 L 为最大聚合损失，u 为初始资金并且满足 $P(L>u)=0.05$，试确定 u。

5. 设 X 为某个正实数点的退化分布：$P(X=x_0)=1$，$x_0>0$，试给出式（7.3.10）中 L_1 的概率密度函数。

6. 设盈余过程中的理赔过程为复合泊松过程，理赔额变量的概率密度函数为 $p(x)=\frac{1}{2}e^{-x}+\frac{3}{4}e^{-3x}+e^{-4x}$，又设调节系数 R 满足方程：

$$1+2R=\frac{1}{2}\left(\frac{1}{1-R}\right)+\frac{1}{4}\left(\frac{3}{3-R}\right)+\frac{1}{4}\left(\frac{4}{4-R}\right)$$

试确定安全附加费率 θ。

7. 对于某复合泊松理赔过程，设理赔额服从混合指数分布，其概率密度函数为 $p(x)=\frac{3}{2}e^{-3x}+\frac{7}{2}e^{-7x}, x>0$，又设附加费率 $\theta=2/5$，试计算破产概率 $\psi(u)$。

8. 若破产概率为 $\psi(u)=0.3e^{-2u}+0.2e^{-4u}+0.1e^{-7u}$，$u\geq 0$，试确定 θ 和 R。

9. 现有两个独立的保险标的，它们都用复合泊松过程来描述，其中理赔额都服从指数分布且平均理赔额为 10。又设保险人 A 和保险人 B 分别承保了这两个保险标的，承保情况如表 7－7 所示。假定保险人 A 与 B 进行了合并，但继续以同样的费率承保了两个标的，试计算其破产概率。

表 7－7

保险人	平均理赔次数	安全附加费率	盈余
A	8	0.4	10
B	2	0.6	5

10. 保险人 A 和 B 共同承保某种保险业务并分别按比例 α 和 $1-\alpha$ 支付每次理赔。设聚合理赔是一个复合泊松过程，理赔的密度函数记做 $f(x)$，x

>0。又设保险人 A 和 B 收取年保费的规则是使得调节系数为固定值 R_A 和 R_B，试证明使总保费最低的 α 为 $R_B/(R_A+R_B)$。

11. 计算泊松盈余过程的最大聚合损失 L 的数学期望 $E(L)$ 和方差 $Var(L)$。

12. 在泊松盈余过程中：泊松参数为 4、理赔额均为 5，单位时间收取的保费 $c=25$，试计算：(1) θ 和 $\psi(0)$；(2) 随机变量 N、L_1 及 L 的数学期望和方差。

13. 已知某离散时间盈余过程的条件如下：$u=2$，年保费为 2.5，年总损失（年底发生）为独立同分布的随机变量，损失量为 0，2，4 的概率分别为 0.5，0.3 和 0.2。计算 $\psi(2, 2)$。

14. 设有某复合泊松理赔过程，其泊松参数为 λ、理赔额变量 C 服从均值为 1 的指数分布，附加费率记作 θ，又记第 n 次理赔或 n 次以前发生破产的概率为 $\psi(u, n)$。(1) 试证明 $\psi(u, 1)=\dfrac{e^{-u}}{2+\theta}$；(2) 将 $\psi(u, 2)$ 表示为 θ 和 u 的函数。

15. 保单理赔总额是一个复合泊松过程，具体数据如下：(1) 通过核保程序的保单数目服从强度为每天 1 000 份的泊松过程；(2) 对于每份通过核保的保单，其被保险人有 20% 的可能性是吸烟者，80% 的可能性是非吸烟者；(3) 保单的定价准则使得每份保单的期望损失是 −100；(4) 对于一个吸烟的被保险人，其理赔额方差为 5 000，而对于一个不吸烟的被保险人，这个数字是 8 000。计算一天内通过核保的所有保单理赔额的方差。

16. 试证明式 (7.7.21)。

17. 某保险公司用复合泊松过程来近似其盈余过程，已知泊松参数为 3.5，理赔额变量的密度函数为 $f(x)=\dfrac{27}{2}x^2e^{-3x}$，$x>0$，单位：百万元。如果保险公司不投入任何初始资本，其破产概率为 80%，如果公司希望 50 年之内不发生破产，那么至少应该投入多少初始资本？

18. 假设 $U(n)=u+cn-S(n)$，$S(n)=W_1+W_2+\cdots+W_n$ 为时间 $[0, n]$ 上的总索赔额。初始盈余为 1 000 元，每个阶段的保费收入为 30 元。除理赔支出外，每个阶段保险人还需支出 8 元的费用。已知 W_1，W_2，…，W_n 独立，服从 $N(10, 4)$ 分布。计算盈余过程的调节系数。

19. 假设 $U(n)=1+1.5n-S(n)$，$S(n)=W_1+W_2+\cdots+W_n$ 为时间 $[0, n]$ 上的总索赔额。对 W_i 的分布如下：$P(W_i=0)=0.40$，$P(W_i=1)=0.30$，$P(W_i=2)=0.20$，$P(W_i=3)=0.10$，$\psi(1, 2)$ 为初始盈余为 1 的情况下到时刻 2 为止的破产概率。计算 $\psi(1, 2)$。

20. 某复合泊松盈余过程为 $U(t)=u+ct-S(t)$，理赔额变量由期望分别为 0.5 和 0.25 的两个指数分布混合而成，权重分别为 75% 和 25%。给定 $\theta=19/21$。计算：(1) $E[\max_{t\geq 0}\{S(t)-ct\}]$；(2) 破产概率 $\psi(2)$。

第二篇　模型的估计和选择

第八章　经 验 模 型

学习目标

- □ 了解完整数据、非完整数据、分组数据的特点，熟悉不同数据情况下的风险集与经验生存函数的计算
- □ 掌握非完整数据生存函数的 Kaplan－Meier 乘积极限估计、死亡力函数的 Nelson-Åalen 估计
- □ 了解生存函数方差估计的基本原理以及对数转换的置信区间的基本原理，掌握生存函数区间估计、Greenwood 方差近似及相应的区间估计
- □ 了解一般核函数估计的理论，熟悉大样本的 Kaplan－Meier 近似计算方法，熟悉多元终止概率的计算，掌握三种常见核函数的密度估计计算

§8.1　数 据 类 型

根据数据完整程度的不同，精算实务中的数据集可以分为完整数据与非完整数据。

8.1.1　完整数据

如果能够对概率分布的任意点收集数据，并且能够记录每个观测值，这种情况称做**完整数据**（complete data）。例如，在生存分析中，当观察者对被观察对象在每个足够小的时间单位内进行观察，直至被观察对象全部死亡，并且除死亡以外，被观察对象不允许离开，则得到的生存数据就是完整的样本数据。例如，当免赔和保单限额都不存在时，保险人对保单损失全额支付，这时获得的保单理赔数据是完整的样本数据。理想的状况是得到每个观测值本质上的精确值，这种情况称为**完整个体数据**（complete individual data）。当个体数据量过多，可以考虑对观测值进行分组，只记录观测值所属的分组，则得到的数据称为**分组数据**。

【例 8－1】　（完整个体数据）这是一组人造的车险赔付数据。假设赔

付按照实际损失全额支付。表 8－1 列出了 20 个赔付额数据样本。

表 8－1　　车险赔付数据

448	2 482	2 753	3 786	3 866	3 965	6 315	6 664	6 707	7 947
8 391	14 540	15 445	15 597	15 602	16 020	22 125	23 879	55 122	220 876

■

【例 8－2】　（完整个体数据）这是一组人造的 10 年定期寿险保单终止时间的数据集。保单的终止是指被保险人身故，或者是被保险人退保（保单合约解除），或者是 10 年保单满期。表 8－2 列出了从发行之日起观测的 30 份保单的数据，其中身故时间中的“—”表示在观察期内没有死亡，退保时间中的“—”表示观察期内没有退保。表格中不仅详细给出了每个投保人的身故时间，还给出了退保时间（只要身故或退保发生在 10 年保单满期之前）。在这个数据集中，无论投保人是否中途退保，他的死亡时间都是能够被观测的，因此属于完整数据。当然，通常情况下我们无法得到已退保的投保人的确切身故时间，也无法得知某个事实上已经身故的投保人如果没有身故将在何时退保。而表 8－2 最后 10 个投保人在 10 年内既未身故又未退保，保单一直持有到满期。

表 8－2　　10 年定期寿险保单终止时间

保单持有人编号	身故时间（年）	退保时间（年）	保单持有人编号	身故时间（年）	退保时间（年）
1	—	0.2	12	5.8	9.2
2	0.5	—	13	—	6.8
3	9.6	1.0	14	6.9	9.2
4	—	1.6	15	—	6.9
5	1.6	7.8	16	—	7.2
6	2.4	5.4	17	8.0	—
7	6.2	3.6	18	8.5	—
8	—	1.8	19	—	8.8
9	—	4.2	20	—	9.5
10	—	5.0	21～30	—	—
11	—	5.6			

■

【例 8－3】　（分组数据）表 8－3 是一组普通责任保险保单的 227 例赔案的赔付额。

【例 8－4】　表 8－4 的数据集搜集了 1956—1958 年间 94 935 个驾驶员

每人每年出现的交通事故数①。注意数据集中“5 次以上”是一个分组。

表 8－3　普通责任保险保单赔付额

赔付额范围（元）	赔付笔数
0～7 500	99
7 500～17 500	42
17 500～32 500	29
32 500～67 500	28
67 500～125 000	17
125 000～300 000	9
300 000 元以上	3

表 8－4　交通事故发生数

赔付次数	赔付笔数
0	81 714
1	11 306
2	1 618
3	250
4	40
5 次以上	7
合计	94 935

■

8.1.2　非完整数据

在实际中经常得到非完整数据。例如，如果保单对任何事故的赔付额都不超过 10 万元，则赔付额数据本身将无从考察那些事故损失量超过 10 万元的保单的实际损失。另一方面，对于 250 元免赔额的车险保单，保险公司不会过问损失额低于 250 元的汽车损伤情况，也不会记录它们的具体数据。因此，实际工作中得到的保单损失数据大多是非完整的。

非完整数据产生的原因有两种：删失或截断。通常情况下我们对属于某一范围内的数据记录其精确值，如果对超出该范围的数据只记录其所属的范围，则得到的数据称为**删失数据**（censored data）；如果对超出该范围的数据不作记录，则得到的数据称为**截断数据**（truncated data）。**左删失数据**（left censored data）是指只知道观测值在某个给定值之下而不知道其具体值；**右删失数据**（right censored data）是指只知道观测值在某个给定值之上同样也不知道其具体值。**左截断数据**（left truncated data）是指对低于某个给定值的观测值不作记录；**右截断数据**（right truncated data）是指对高于某个给定值的观测值不作记录。

最常见的非完整数据是左截断和右删失数据。例如，存在免赔额的非寿险保单赔付数据是左截断数据；存在赔偿限额的非寿险保单赔付数据是右删失数据。在生命表构造中，要跟踪每个人的出生和身故是不现实的，因而常见的做法是在几年的时间内，跟踪观测一个由不同年龄段的人组成的人群的生存状况。当某人开始参与一项生存研究项目时，他必然处于生

① 数据来源：Dropkin, L. “Some Considerations on Automobile Rating System Utilizing Individual Driving Records”, Proc. of the Casualty Actuarial Society, XLVI, 1959, 391－405.

存状态，且其身故年龄至少不会低于他参加该项目时的年龄，因此数据是左截断的。如果这个参与者在退出该研究项目时还未身故，就出现了右删失的现象，具体身故年龄无法确定，但是只能确定身故年龄不会小于此人在退出该项目时的年龄。

【例 8－5】　（删失数据）例 8－2 中的数据仅是理论上存在的，实际中我们只能观测到某个事件（死亡、退保）的首次发生时间。我们在此基础上新增 10 位保单持有人的观测值，这 10 名保单持有人是从保单发行一段时间后才开始观测的。“初始观测时间”给出了每张保单的开始观测时间；“最后观测时间”给出了每个保单终止的时间；“事件”一栏中记录了发生事件的属性，s 表示退保，d 表示身故，e 表示满期。得到的数据见表 8－5。这实际上是一组删失和截断都存在的数据。对于编号 31～40 的保单持有人，其死亡时间晚于初始观测时间，因此是左截断的。另一方面，对于退保以及满期保单的持有人，只能知道其生存时间大于某一值，而不能确切知道其生存年数，因此是右删失的。

表 8－5　　　　10 年定期保单观测数据集

编号	初始观测时间	最后观测时间	事件	编号	初始观测时间	最后观测时间	事件
1	0	0.2	s	17	0	8.0	d
2	0	0.5	d	18	0	8.5	d
3	0	1.0	s	19	0	8.8	s
4	0	1.6	s	20	0	9.5	s
5	0	1.6	d	21～30	0	10.0	e
6	0	2.4	d	31	0.6	10.0	e
7	0	3.6	s	32	1.4	10.0	e
8	0	3.8	s	33	2.0	8.2	d
9	0	4.2	s	34	3.6	6.2	d
10	0	5.0	s	35	4.2	7.8	s
11	0	5.6	s	36	5.8	9.6	s
12	0	5.8	d	37	6.0	8.0	d
13	0	6.8	s	38	6.4	10.0	e
14	0	6.9	d	39	7.8	10.0	e
15	0	6.9	s	40	7.9	10.0	e
16	0	7.2	s				

■

【例 8－6】　（删失数据）表 8－6 是以日历年 2009 年为观测期间得到的 5 人组成的样本。对每个样本记录其生日和死亡日期。对于 2009 年内没有死亡的个体，我们无法得到其死亡时间，因此得到的数据是删失的。

表 8 – 6　　生存分析中的样本数据

观察者	生日	死亡日期
1	1979.7.1	—
2	1979.4.1	—
3	1979.1.1	2009.10.1
4	1979.7.1	—
5	1978.4.1	2009.4.22

■

§8.2 完整数据情况下的经验分布函数估计

8.2.1 个体数据

1. 经验分布函数。为了引入经验分布函数，我们首先引入**数据依赖型分布**（data dependent distribution）的概念。数据依赖型分布是一种非参数分布，它的复杂程度至少与产生它的数据或者其他信息相当，并且其参数个数会随着数据点或者信息量的增加而增加。数据依赖型分布中，最简单的分布是**经验分布**（empirical distribution）。为研究某总体分布函数，我们从中抽取一个样本量为 n 的数据集，如果假设每个数据点的概率为 $1/n$，则得到的分布函数即为经验分布，其计算公式为：

$$F_n(x) = \sum_{i=1}^{n} \frac{1}{n} I(x_i \leqslant x) = \frac{\#\{x_i \leqslant x\}}{n} \tag{8.2.1}$$

其中：$I(\cdot)$ 为示性函数，$\#A$ 表示集合 A 中元素的个数。

观测值 x_i 中可能包括相同的数值；我们记其中有 k 个不同的数值，$y_1 < y_2 < \cdots < y_k$，并且记 $s_j = \#\{x_i: x_i - y_j\}$ 表示取值 y_j 的观测值的个数。容易知道：$s_1 + s_2 + \cdots + s_k = n$。则经验分布又可以写成：

$$F_n(x) = \frac{1}{n} \sum_{j:y_j \leqslant x} s_j \tag{8.2.2}$$

经验分布概率函数（empirical distribution probability function）$p_n(x)$ 定义为取值 x 的观测值的个数。显然，

$$p_n(x) = \begin{cases} \dfrac{s_j}{n}, & x = y_j \\ 0, & x = \text{其他} \end{cases} \tag{8.2.3}$$

并且 $p_n(x)$ 和 $F_n(x)$ 的关系为：

$$F_n(x) = \sum_{j:y_j \leqslant x} p_n(y_j) \tag{8.2.4}$$

注意式（8.2.2）和（8.2.4）是一致的。

【例 8 – 7】 假设样本数据 $x_1, x_2, \cdots, x_8$ 为 7、2、4、4、6、2、1、

9。求在各个观测值上的经验概率。

解： 将数据排序，可以得到6个不同的观测值：$y_1=1$，$y_2=2$，$y_3=4$，$y_4=6$，$y_5=7$，$y_6=9$。

相应地观测个数为：$s_1=1$，$s_2=2$，$s_3=2$，$s_4=1$，$s_5=1$，$s_6=1$。从而有：

$$p_8(1)=\frac{1}{8},\ p_8(2)=\frac{1}{4},\ p_8(4)=\frac{1}{4},\ p_8(6)=\frac{1}{8},\ p_8(7)=\frac{1}{8},\ p_8(9)=\frac{1}{8}$$

■

【例8-8】 利用例8-1中的数据绘制经验分布函数图，并计算$F_{20}(10\ 000)$。

解： $\because \#\{x_i\leqslant 10\ 000\}=11$，$n=20$　$\therefore F_{20}(10\ 000)=\dfrac{11}{20}=0.55$

绘制的经验分布函数图见图8-1。从图8-1中可以看出，经验分布函数实际上是一个在每个数据点跳跃的阶梯函数，在观测值y_j处跳跃值为$p_n(y_j)=s_j/n$。■

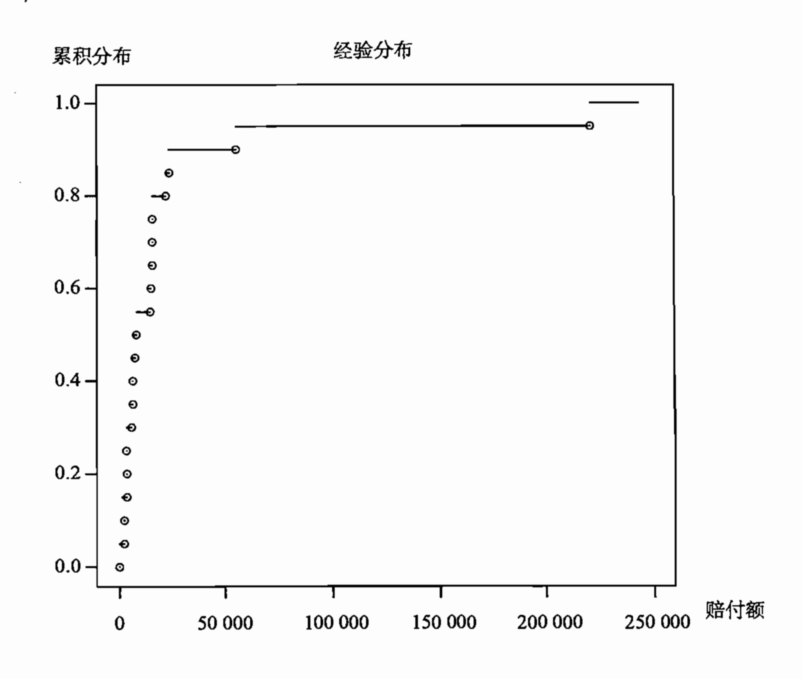


图8-1　例8-1的经验分布函数

经验分布函数作为分段常数函数的这一性质对总体分布的推断意义不是很大，因为总体分布函数几乎总是严格递增的。事实上，经验分布函数的阶梯式增加是由于样本容量的有限性；可以推断，如果样本容量无限增大，经验分布函数阶梯的“个数”也就不断增大，而“幅度”不断减小，因此得到的曲线也会更加平滑。因此实际工作中可以用拟合这些降幅的光

滑曲线来调整图 8－1 中的曲线。这一调整过程称为修匀，它是推导总体分布函数 $F(x)$ 程序的一部分。在此，我们只考察 $F(x)$ 的初始估计量程序而不进行修匀。

2. 风险集与经验生存函数。和经验分布函数对应，**经验生存函数**（empirical survival function）$S_n(x)$ 指的是对生存时间大于 x 的概率的估计，即大于 x 的观测值占总体数据集的比例。经验生存函数的定义是：

$$S_n(x)=1-F_n(x) \tag{8.2.5}$$

由式（8.2.1）和（8.2.5）知：

$$S_n(x)=1-F_n(x)=1-\frac{\#\{x_i\leqslant x\}}{n}=\frac{\#\{x_i>x\}}{n}$$

在观测值 y_j 处的**风险集**（risk set）是指不小于 y_j 的观测值组成的集合，记做：

$$r_j=\{x_i:x_i\geqslant y_j\} \tag{8.2.6}$$

在生存分析中，由于观测值为个体的存活时间，因此观测值 y_j 处风险集是指在时刻 y_j 面临死亡风险的被观测者的全体。这些被观测者在时刻 y_j 死亡，或者以后在 y_j 死亡。它们构成时刻 y_j 死亡事件的风险暴露组。在理赔额模型中，观测值为理赔额的数值，理赔额 y_j 处的风险集表示理赔额可能大于等于 y_j 的观测对象的集合。

在计算中，我们也用 r_j 表示观测值 y_j 处的风险集的大小，即

$$r_j=\#\ \{x_i:\ x_i\geqslant y_j\} \tag{8.2.7}$$

由式（8.2.2）知：

$$r_j=\sum_{i=j}^{k}s_i \tag{8.2.8}$$

由式（8.2.2）、(8.2.5）和（8.2.8）知，经验生存函数也可如下计算：

$$S_n(x)=\frac{r_j}{n},y_{j-1}\leqslant x<y_j \tag{8.2.9}$$

相应的经验分布函数为：

$$F_n(x)=1-\frac{r_j}{n},y_{j-1}\leqslant x<y_j \tag{8.2.10}$$

【例 8－9】 利用例 8－7 的数据，计算各个观测值处的风险集大小，并绘制经验生存函数的图形。

解： 由式（8.2.8），将 $s_1=1$，$s_2=2$，$s_3=2$，$s_4=1$，$s_5=1$，$s_6=1$ 代入得：

$r_1=\sum_{i=1}^{6}s_i=8$，$r_2=\sum_{i=2}^{6}s_i=7$，同理得：$r_3=5$，$r_4=3$，$r_5=2$，$r_6=1$。

经验生存函数图见图 8－2。

3. 经验分布函数的统计分析。经验分布函数 $F_n(x)$ 是对总体分布 $F(x)$ 函数的一种估计量，下面研究它的统计性质，首先看它的相合性。由式

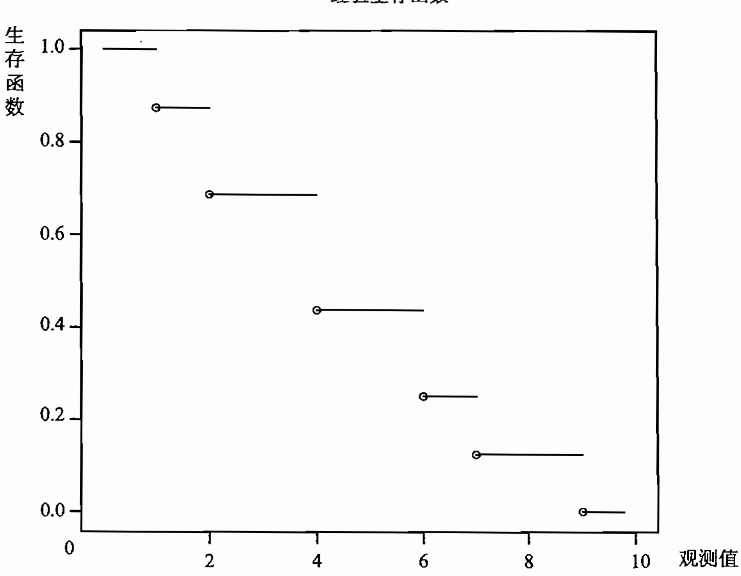


图 8-2　例 8-9 中的经验生存函数

■

(8.2.1)：

$$F_n(x) = \sum_{i=1}^{n} \frac{1}{n} I(x_i \leqslant x) = \frac{\#\{x_i \leqslant x\}}{n}$$

其中每一个加项都独立并且服从两点分布 $I(x_i \leqslant x) \sim B(F(x))$，因此 $\#\{x_i \leqslant x\}$ 服从二项分布 $B(n, F(x))$：因此，

$$E[I(x_i \leqslant x)] = F(x) \tag{8.2.11}$$

$$Var[I(x_i \leqslant x)] = F(x)[1 - F(x)] \tag{8.2.12}$$

将式（8.2.11）、（8.2.12），代入式（8.2.1），得到：

$$E[F_n(x)] = \sum_{i=1}^{n} \frac{1}{n} E[I(x_i \leqslant x)] = \sum_{i=1}^{n} \frac{1}{n} F(x) = F(x) \tag{8.2.13}$$

$$Var[F_n(x)] = Var[\sum_{i=1}^{n} \frac{1}{n} I(x_i \leqslant x)] = \frac{1}{n^2} \sum_{i=1}^{n} Var[I(x_i \leqslant x)]$$

$$= \frac{1}{n} F(x)[1 - F(x)]$$

即

$$Var[F_n(x)] = \frac{1}{n} F(x) S(x) \tag{8.2.14}$$

因此，$F_n(x) \xrightarrow{P} F(x)$，经验分布函数 $F_n(x)$ 是对总体分布 $F(x)$ 函数的无偏一致估计量。

实际计算中，方差的估计可如下计算：

$$\widehat{Var}[F_n(x)] = \frac{1}{n} F_n(x) S_n(x) \tag{8.2.15}$$

类似地，可以得到经验生存函数的相合性：

$$E[S_n(x)]=S(x)$$

$$Var[S_n(x)]=\frac{1}{n}F(x)S(x) \tag{8.2.16}$$

以及方差的估计：

$$\widehat{Var}[S_n(x)]=\frac{1}{n}F_n(x)S_n(x) \tag{8.2.17}$$

在使用经验估计时，有时我们关心概率 $p=P\{a<X\leqslant b\}$ 的估计。此时 p 的经验估计为 $\hat{p}=F_n(b)-F_n(a)$。与上文的推导类似，可以知道 $E(\hat{p})=p$，$Var(\hat{p})=\frac{p(1-p)}{n}$，即估计量也是相合的。

对于一般的经验估计 $\hat{p}$，由中心极限定理，$\hat{p}$ 近似地服从正态分布：

$$\frac{\hat{p}-p}{\sqrt{Var(\hat{p})}}=\frac{(\hat{p}-p)}{\sqrt{p(1-p)/n}}\xrightarrow{d}N(0,\ 1)$$

因此 p 的 $100(1-\alpha)\%$ 的置信区间可通过求解如下不等式得到：

$$-u_{\alpha/2}\leqslant\frac{(\hat{p}-p)}{\sqrt{p\ (1-p)\ /n}}\leqslant u_{\alpha/2} \tag{8.2.18}$$

其中 $u_{\alpha/2}$ 是标准正态分布的 $100(1-\alpha/2)\%$ 分位点。有时这样做是很困难的（因为 p 出现在分母中），所以如果必要时可以将式（8.2.18）分母中的 p 用 $\hat{p}$ 代替，从而得到如下近似置信区间：

$$\left[\hat{p}-u_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+u_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{8.2.19}$$

【例 8－10】 利用例 8－2 中的数据，用经验方法估计 ${}_{2|}q$ 和 q_2 并计算估计量的方差。

解： 注意 ${}_{2|}q$ 和 q_2 是不同的，${}_{2|}q$ 是个体在第三年死亡的概率，而 q_2 是个体在活过两年的情况下在第三年死亡的概率。如果用 M 表示在 0～2 之间身故的人数，N 表示在 2～3 之间身故的人数，则可以得到：${}_{2|}\hat{q}=\frac{N}{n}$，$\hat{q}_2=\frac{N}{n-M}$。数据集中的 $n=30$，因此，

$${}_{2|}\hat{q}=S_{30}(2)-S_{30}(3)=\frac{1}{30},\ Var({}_{2|}\hat{q})=\frac{{}_{2|}\hat{q}(1-{}_{2|}\hat{q})}{n}=\frac{29}{30^3},\ \hat{q}_2=\frac{1}{30-2}=\frac{1}{28}$$

对于 $\hat{q}_2$ 的方差 $Var(\hat{q}_2)=Var\left(\frac{N}{n-M}\right)$ 通常是无法求得的，因为 M 取值为 n 的概率为正数。通常的做法是求取条件方差，即给定 $M=2$ 的条件下的方差，此时有：

$$\widehat{Var}(\hat{q}_2\mid S_{30}(2)=\frac{28}{30})=\frac{(1/28)(27/28)}{28}=0.00123$$ ■

例8－10的结果可以推广到一般情形。假设最初样本量为n，n_x为时刻x存活的个体数，n_{x+n}为时刻$x+n$存活的个体数，则${}_nq_x$表示x岁生存的人在$x+n$岁前死亡的概率，则

$${}_nq_x = P(x < X \leqslant x+n \mid x > x) = \frac{S(x) - S(x+n)}{S(x)}$$

因此${}_nq_x$的估计值为：

$${}_n\hat{q}_x = \frac{n_x - n_{x+n}}{n_x}$$

${}_n\hat{q}_x$的条件方差的估计为：

$$\widehat{Var}\left[{}_n\hat{q}_x \mid \text{在} x \text{时刻有} n_x \text{个人}\right] = {}_n\hat{q}_x \frac{1 - {}_n\hat{q}_x}{n_x} = \frac{(n_x - n_{x+n})n_{x+n}}{n_x^3}$$

【例8－11】　计算例8－10中${}_{21}q$的95%置信区间。

解：将$u_{\alpha/2}=1.96$，${}_{21}\hat{q}=\frac{1}{30}$，$n=30$代入式（8.2.18）中，解得95%置信区间为$0.0059 < {}_{21}q < 0.1667$。若代入式（8.2.19），得到近似的区间为$(-0.0309,\ 0.0976)$。■

8.2.2　分组数据

1. 卵形线和直方图。对于分组数据，根据之前的定义构造经验分布函数是不可能的。但是，仍然可以近似估计经验分布函数，方法是利用插值方法估计，其中最常见的方法是线性插值。

假设分组数据的分界点为$c_0 < c_1 < \cdots < c_k$，其中$c_0=0$，$c_k=\infty$，落在c_{j-1}与c_j之间的观测值有n_j个，则有：$\sum_{j=1}^{k} n_j = n$。对这样的数据，经验分布函数在分界点处的值为：

$$F_n(c_j) = \sum_{i=1}^{j} n_i/n \tag{8.2.20}$$

注意：要使$F_n(c_j)$成为经验分布函数，由分布函数的右连续性，应当保证分组数据包括左端点c_{j-1}不包括右端点c_j。

为了计算分布函数在任意观测处的取值，可以采用线性插值法，得到的估计值叫做**经验分布光滑曲线**，简称**卵形线**（ogive）。计算公式为：

$$F_n(x) = \frac{c_j - x}{c_j - c_{j-1}}F(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}}F(c_j),\qquad c_{j-1} \leqslant x < c_j \tag{8.2.21}$$

该函数是几乎处处可导的，因此可以人为地将其密度函数定义为右连续的函数即可，得到的密度函数称为**直方图**（histogram）。计算公式为：

$$f_n(x) = \frac{n_j}{n(c_j - c_{j-1})},\qquad c_{j-1} \leqslant x < c_j \tag{8.2.22}$$

注意，这里定义的直方图，其组距可以是不同的。很多计算机程序生

成的直方图其实只是柱状图，柱的高度和频数成比例。在区间长度相同的情况下，这种方法是可以接受的，但是当区间长度不同时，就必须使用公式（8.2.22）进行计算。这种方法的优点是：直方图确实代表密度函数，而且直方图下面的面积可以代表经验概率值。

【例 8－12】 利用例 8－3 中的数据，构造卵形线绘制直方图。

解： 分布函数为：

$$F_{227}(x)=\begin{cases}0.000058150x, & 0\leqslant x<7500\\ 0.29736+0.000018502x, & 7\,500\leqslant x<17\,500\\ 0.47210+0.000008517x, & 17\,500\leqslant x<32\,500\\ 0.63436+0.000003524x, & 32\,500\leqslant x<67\,500\\ 0.78433+0.000001302x, & 67\,500\leqslant x<125\,000\\ 0.91882+0.000000227x, & 125\,000\leqslant x<300\,000\\ \text{未定义}, & x\geqslant 300\,000\end{cases}$$

由于卵形线是分段线性的，其导数的计算是显然的，此处从略。相应的直方图见图 8－3。■

【例 8－13】 考察一个在 $t=0$ 处有 20 个个体的样本，所有个体均在 5 周内死亡，并且只记录每周死亡的人数。在 5 周内死亡的人数分别为：2，3，8，6，1。运用所给数据估计$\hat{q}_3$，并绘制直方图。

解： q_3 是在第 3 周末存活的人在第 4 周死亡的概率。因此$\hat{q}_3=\dfrac{6}{7}$。相应的直方图见图 8－4。

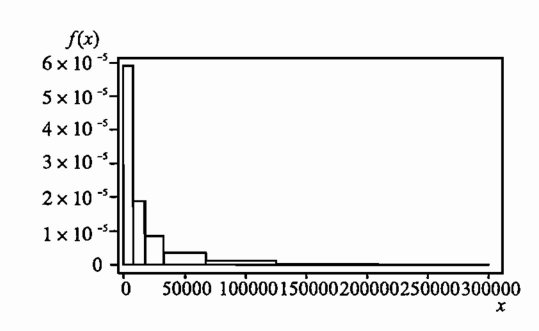


图 8－3 例 8－13 中赔付额分布的直方图

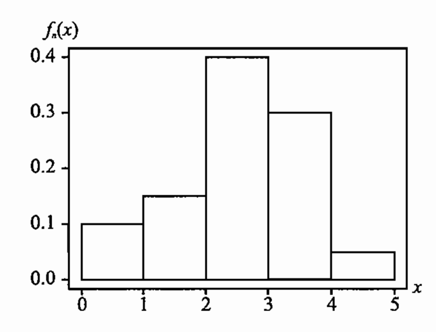


图 8－4 例 8－14 死亡时间分布的直方图

■

2. 密度估计的性质。设 N_j 表示落在 c_{j-1}和 c_j 之间的观测值个数。$N_1,N_2,\cdots,N_k$ 服从联合多项分布，其联合概率函数为：

$$p(n_1,n_2,\cdots,n_k)=\frac{n!}{n_1!\ n_2!\ \cdots n_k!}\prod_{j=1}^{k}(F(c_j)-F(c_{j-1}))^{n_j}\qquad(8.2.23)$$

从而其边际分布为二项分布：

$$N_j\sim B(n,F(c_j)-F(c_{j-1}))\qquad(8.2.24)$$

在时刻 c_{j-1} 前死亡的总人数的分布为：

$$\sum_{i=1}^{j-1} N_i \sim B(n, F(c_{j-1})) \tag{8.2.25}$$

因此，

$$E(N_j)=n(F(c_j)-F(c_{j-1})), Var(N_j)=n(F(c_j)-F(c_{j-1}))(1-F(c_j)+F(c_{j-1}))$$

$$E\left(\sum_{i=1}^{j-1} N_i\right) = nF(c_{j-1}),\ Var\left(\sum_{i=1}^{j-1} N_i\right) = n(1-F(c_{j-1}))F(c_{j-1})$$

由式（8.2.21）知，对于 $c_{j-1} \leqslant x < c_j$

$$F_n(x) = F_n(c_{j-1}) + \frac{F_n(c_j)-F_n(c_{j-1})}{c_j - c_{j-1}}(x - c_{j-1})$$

$$= \frac{1}{n}\sum_{i=1}^{j-1} n_i + \frac{n_j}{n} \cdot \frac{(x-c_{j-1})}{c_j - c_{j-1}}$$

将以上结果代入得：

$$E(F_n(x)) = \frac{1}{n}nF(c_{j-1}) + \frac{n[F(c_j)-F(c_{j-1})]}{n} \cdot \frac{x-c_{j-1}}{(c_j-c_{j-1})}$$

$$= F(c_{j-1})\frac{c_j - x}{c_j - c_{j-1}} + F(c_j)\frac{x-c_{j-1}}{c_j-c_{j-1}} \tag{8.2.26}$$

当 $x \neq c_{j-1}$ 时，这个估计是有偏的。其方差计算比较复杂：

$$Var(F_n(x)) = \frac{1}{n^2}Var\left(\sum_{i=1}^{j-1} N_i\right) + \frac{(x-c_{j-1})^2}{n^2(c_j-c_{j-1})^2}Var(N_j) + \frac{2(x-c_{j-1})}{n^2(c_j-c_{j-1})}Cov\left(\sum_{i=1}^{j-1} N_i, N_j\right)$$

$$= \frac{1}{n^2}\{nF(c_{j-1})[1-F(c_{j-1})]\} + \frac{(x-c_{j-1})^2}{n^2(c_j-c_{j-1})^2}\{n[F(c_j)-F(c_{j-1})]$$

$$\cdot[1-F(c_j)+F(c_{j-1})]\} + \frac{2(x-c_{j-1})}{n^2(c_j-c_{j-1})}\{-nF(c_{j-1})[F(c_j)-F(c_{j-1})]\}$$

$$= \frac{1}{n}\{F(c_{j-1}) + \left(\frac{x-c_{j-1}}{c_j-c_{j-1}}\right)^2[F(c_j)-F(c_{j-1})] - [F(c_{j-1}) + \frac{x-c_{j-1}}{c_j-c_{j-1}}(F(c_j)$$

$$-F(c_{j-1}))]^2\}$$

$$= \frac{1}{n}\{F(c_{j-1}) + (\frac{x-c_{j-1}}{c_j-c_{j-1}})^2[F(c_j)-F(c_{j-1})] - [E(F_n(x))]^2\} \tag{8.2.27}$$

类似地，代入式（8.2.22），得到：

$$E[f_n(x)] = \frac{F(c_j)-F(c_{j-1})}{c_j-c_{j-1}}, c_{j-1} \leqslant x < c_j \tag{8.2.28}$$

$$Var[f_n(x)] = \frac{(F(c_j)-F(c_{j-1}))(1-F(c_j)+F(c_{j-1}))}{n(c_j-c_{j-1})^2} \tag{8.2.29}$$

$Var[f_n(x)]$ 的估计值为：

$$\widehat{Var}[f_n(x)] = \frac{f_n(x)(1-f_n(x)(c_j-c_{j-1}))}{n(c_j-c_{j-1})},\ c_{j-1} \leqslant x < c_j$$

这些结果和完整个体数据的估计都是基于二项分布进行的，因此结果也是一致的。

【例 8－14】 在例 8－13 中的样本，已知 $n_4+n_5=7$ 表示第三周末的存活人数是 7。

(1) 假定样本的生存分布为 (0，5] 上的均匀分布，求 $Var({}_{21}\hat{q})$ 和 $Var(\hat{p}_3 \mid n_4+n_5=7)$

(2) 若生存分布未知，求 $Var({}_{21}\hat{q}_0)$ 和 $Var(\hat{p}_3 \mid n_4+n_5)$ 的估计。

解：(1) 若生存分布为均匀分布，则 ${}_{21}q_0=F(3)-F(2)=0.2$，$p_3=\dfrac{S(4)}{S(3)}=0.5$。从而

$$Var({}_{21}\hat{q}_0)=\frac{{}_{21}q_0(1-{}_{21}q_0)}{n}=\frac{0.2\times0.8}{20}=0.008$$

$$Var(\hat{p}_3 \mid n_4+n_5)=\frac{p_3(1-p_3)}{n_4+n_5}=\frac{0.5\times0.5}{7}=0.03571$$

(2) 以 ${}_{21}\hat{q}_0=0.4$，$\hat{p}_3=\dfrac{1}{7}$ 代入方差估计式，得到：

$$\widehat{Var}({}_{21}\hat{q}_0)={}_{21}\hat{q}_0(1-{}_{21}\hat{q}_0)/n=0.012,$$

$$\widehat{Var}(\hat{p}_3 \mid n_4+n_5=7)=\hat{p}_3(1-\hat{p}_3)/(n_4+n_5)=0.01749。$$ ■

【例 8－15】 利用例 8－4 的数据，计算 $f_n(1)$ 的方差。

解：利用式 (8.2.30) 计算，其中 $n=94\ 935$，$c_2=2$，$c_1=1$，

$$f_n(1)=\frac{11\ 306}{94\ 935\times(2-1)}=0.11909。$$ 得到：

$$Var(f_n(1))=\frac{f_n(1)(1-f_n(1)(2-1))}{94\ 935\times(2-1)}=1.10506\times10^{-6}$$ ■

§8.3 非完整数据情况下的经验分布函数估计

8.3.1 风险集

为了用删失或者截断的数据来构造经验分布函数，第一个任务是明确一些记号来表示与数据信息相关的量。对于个体数据，如下两个因素是必须考虑的：

1. 数据观测值的截断点，用 d_j 表示，如果没有截断发生，则 $d_j=0$。

2. 数据观测值本身。在理论研究中，对于每一个观测对象，都假定存在一个事件发生点 x_j 和删失值 u_j，但是能够被观测的只能是二者中的较小值；因此，如果该数据是删失值，将其值记为 u_j，否则记为 x_j。

和完整数据的情形类似，把未被删失的观测值 x_i 中包括的所有 k 个不同的数值记做 $y_1<y_2<\cdots<y_k$，并且用 s_j 表示这些 y_j 出现的次数。和完整数据不同的是风险集 r_j 的计算。在生存模型中，**风险集是由那些在指定年龄**

仍处于被观察状态的个体构成。在观测值 y_j 时刻的风险集包括：

（1）死亡时间在 y_j 或 y_j 以后的个体；

（2）删失时间在 y_j 或 y_j 以后的个体。

但是，对于那些在 y_j 以后才首次被观测到的个体，我们认为其在时刻 y_j 并没有处于被观测状态。因此，对于风险集大小的计算公式为：

$$r_j = \#\{x_i : x_i \geqslant y_j\} + \#\{u_i : u_i \geqslant y_j\} - \#\{d_i : d_i \geqslant y_j\} \tag{8.3.1}$$

另外，注意到：

$$n = \#\{x_i\} + \#\{u_i\} = \#\{x_i : x_i \geqslant y_j\} + \#\{x_i : x_i < y_j\} + \#\{u_i : u_i \geqslant y_j\} + \#\{u_i : u_i < y_j\} = \#\{d_i\} = \#\{d_i : d_i \geqslant y_j\} + \#\{d_i : d_i < y_j\}$$

因此，式（8.3.1）也可写做：

$$r_j = \#\{d_i : d_i < y_j\} - \#\{x_i : x_i < y_j\} - \#\{u_i : u_i < y_j\} \tag{8.3.2}$$

在数据没有删失和截断的情况下，容易知道，式（8.2.7）是式（8.3.1）、（8.3.2）的特殊情形。式（8.3.2）从直观意义上更加容易理解，因为它是由所有在给定年龄之前被观测的个体减去已经离去的个体数。此外，实际计算中，如下的递推公式可以使计算更加方便：

$$r_j = r_{j-1} + \#\{d_i : y_{j-1} \leqslant d_i < y_j\} - \#\{x_i : y_{j-1} \leqslant x_i < y_j\} - \#\{u_i : y_{j-1} \leqslant u_i < y_j\} \tag{8.3.3}$$

其中规定 $r_0 = 0$。

在损失模型中，如果观测值是损失额。风险集就是损失观测值（实际金额或者由保单限额决定的最大金额）大于或者等于 y_j 的保单数减去免赔额大于或者等于 y_j 的保单数，计算公式同上。

【例 8－16】　对例 8－5 的数据，同时利用式（8.3.2）或（8.3.3）计算以上定义的各种量。

解：计算结果列于表 8－7 和表 8－8。

表 8－7　　**例 8－16 的数据**

i	d_i	x_i	u_i	i	d_i	x_i	u_i
1	0	—	0.2	17	0	8.0	—
2	0	0.5	—	18	0	8.5	—
3	0	—	1.0	19	0	—	8.8
4	0	—	1.6	20	0	—	9.5
5	0	1.6	—	21～30	0	—	10.0
6	0	2.4	—	31	0.6	—	10.0
7	0	—	3.6	32	1.4	—	10.0
8	0	—	3.8	33	2.0	8.2	—
9	0	—	4.2	34	3.6	6.2	—

续表

i	d_i	x_i	u_i	i	d_i	x_i	u_i
10	0	—	5.0	35	4.2	—	7.8
11	0	—	5.6	36	5.8	—	9.6
12	0	5.8	—	37	6.0	8.0	—
13	0	—	6.8	38	6.4	—	10.0
14	0	6.9	—	39	7.8	—	10.0
15	0	—	6.9	40	7.9	—	10.0
16	0	—	7.2				

表 8-8　　　　例 8-16 的风险集计算

j	y_i	s_j	r_j
1	0.5	1	30-0-1=29 或 0+30-0-1=29
2	1.6	1	32-1-2=29 或 29+2-1-1=29
3	2.4	1	33-2-3=28 或 29+1-1-1=28
4	5.8	1	35-3-8=24 或 28+2-1-5=24
5	6.2	1	37-4-8=25 或 24+2-1-0=25
6	6.9	1	38-5-9=24 或 25+0-1-0=24
7	8.0	2	40-6-12=22 或 24+2-1-3=22
8	8.2	1	40-8-12=20 或 22+0-2-0=20
9	8.5	1	40-9-12=19 或 20+0-1-0=19

■

8.3.2　Kaplan-Meier 乘积极限估计

（一） Kaplan-Meier 估计的推导

在给出了风险集概念的基础上，我们就可以对生存函数作进一步估计。最常用的方法是 Kaplan-Meier 乘积极限估计 （Kaplan-Meier product-limit Estimator），它是由 Kaplan P. 和 Meier E. 在 1958 年提出的。其估计生存函数的方法是：首先，把未被删失的观测值 x_i 中包括的所有 k 个不同的数值记做 $y_1<y_2<\cdots<y_k$。其次，对于 $t<y_1$ 的值，因为在 y_1 之前没有人身故，所以认为 $S(t)=1$。然后，在 y_1 时刻的死亡事件发生以前，风险集为 r_1，它表示可能有 r_1 人会面临死亡风险，而实际死亡的人数为 s_1。因此，活过时刻 y_1 的概率为$\frac{r_1-s_1}{r_1}$，这就是 $S(y_1)$ 的值。在 y_1 和 y_2 之间的时刻，没有新的死亡发生，因此在这个时间段内生存函数值保持不变。而在时刻 y_2，由相同方法知道活过这个时刻的条件概率为 $P\{T>y_2\mid T>y_1\}=\frac{r_2-s_2}{r_2}$，而

我们假设观察对象之间是相互独立的，因此 $S(y_2)=S(y_1)\frac{r_2-s_2}{r_2}$。继续以上推导，就可以得到 Kaplan－Meier 乘积极限公式：

$$S_n(t)=\begin{cases}1, & 0\leqslant t<y_1\\ \prod_{i=1}^{j-1}\frac{r_i-s_i}{r_i}, & y_{j-1}\leqslant t<y_j, j\leqslant k\\ \prod_{i=1}^{k}\frac{r_i-s_i}{r_i}\text{或}0, & t\geqslant y_k\end{cases}\tag{8.3.4}$$

值得注意的是，当 $t\geqslant y_k$ 时 $S_n(t)$ 的值。如果 $s_k=r_k$，则当 $t\geqslant y_k$ 时，该样本中的所有个体都在 y_k 之前身故，因此从经验上看，$S_n(t)=0$。但是由于删失的存在，可能在最后一个身故时刻被观察以后，仍有个体生存，而这些个体在 y_k 以前被删失了，因此在 y_k 以后的生存函数无法估计。对于这种情况有三种方法。第一种解决办法是取最后得到的函数值 $\prod_{i=1}^{k}\frac{r_i-s_i}{r_i}$ 作为 $S(t)$ 的估计，这种方法得到的估计量是有偏的，而且当 $s_k<r_k$ 时，利用生存函数求取分布的矩时所作的广义积分发散，这个性质是不理想的。因此，有时我们也采用第二种办法，即规定在 y_k 以后的所有取值都为 0，这种方法得到的估计量也是有偏的，但是可以用于计算各阶矩。第三种方法是折中的方法，即选择一条指数函数衰减的曲线去拟合 y_k 以后的生存函数。令 $w=\max\{y_k, u_1, u_2, \cdots, u_n\}$，对 $t>w$，规定 $S_n(t)=e^{\ln S^{*}\cdot\frac{t}{w}}=S^{*\frac{t}{w}}$，其中 $s^{*}=\prod_{i=1}^{k}\frac{r_i-s_i}{r_i}$。由指数函数的性质，此时的经验分布函数对各阶矩的积分都是收敛的。

还要注意，在式（8.3.4）中的 $S_n(t)$ 和经验生存函数式（8.2.9）的符号是相同的。这并不会引起混淆，因为在完整数据的情况下，Kaplan－Meier 估计式（8.3.4）正好退化为经验分布函数（8.2.9）。在完整数据的情形，所有观测值均为 x_i，并且 $d_i=0$。因此，$r_1=n$，$r_i=\sum_{j=i}^{k}s_j$，$r_i-s_i=\sum_{j=i}^{k}s_j-s_i=\sum_{j=i+1}^{k}s_j=r_{i+1}$。将这些关系式代入式（8.3.4），有

$$S_n(t)=\begin{cases}1, & 0\leqslant t<y_1\\ \prod_{i=1}^{j-1}\frac{r_{i+1}}{r_i}=\frac{r_j}{n}, & y_{j-1}\leqslant t<y_j, j\leqslant k\\ \prod_{i=1}^{k}\frac{r_{i+1}}{r_i}=\frac{r_k}{n}, & t\geqslant y_k\end{cases}\tag{8.3.5}$$

此处的式（8.3.5）显然和式（8.2.9）是等价的。

【例 8－17】 利用 Kaplan－Meier 方法对例 8－5 中的数据的生存函数

进行估计。

解：利用例 8－16 的结果，代人计算得到：

$$S_{40}(t)=\begin{cases}1, & 0\leqslant t<0.5\\ (29-1)/29=0.9655, & 0.5\leqslant t<1.6\\ 0.9655(29-1)/29=0.9322, & 1.6\leqslant t<2.4\\ 0.9322(28-1)/28=0.8989, & 2.4\leqslant t<5.8\\ 0.8989(24-1)/24=0.8615, & 5.8\leqslant t<6.2\\ 0.8615(25-1)/25=0.8270, & 6.2\leqslant t<6.9\\ 0.8270(24-1)/24=0.7925, & 6.9\leqslant t<8.0\\ 0.7926(22-2)/22=0.7205, & 8.0\leqslant t<8.2\\ 0.7205(20-1)/20=0.6845, & 8.2\leqslant t<8.5\\ 0.6845(19-1)/19=0.6485, & 8.5\leqslant t<10.0\\ 0.6485 \text{ 或 } 0 \text{ 或 } 0.6485^{\frac{t}{10.0}}, & t\geqslant 10.0\end{cases}$$

■

（二）方差估计

Kaplan－Meier 公式本质上分为如下两个步骤：

首先将生存函数分解成一系列条件概率的乘积，再对每个条件概率进行估计，即：

$$S(t)=P\{T>t\}=\prod_{i=1}^{j-1}P\{T>y_i|T>y_{i-1}\},\ y_{j-1}\leqslant t<y_j,\ j\leqslant k \qquad (8.3.6)$$

其次，构造每一个条件概率的无偏估计：

$$\hat{P}\{T>y_i|T>y_{i-1}\}=\frac{r_i-s_i}{r_i} \qquad (8.3.7)$$

由死亡事件的独立性，将式（8.3.7）代人式（8.3.6）便得到估计公式（8.3.4）。

对于式（8.3.7）的估计，注意到 r_i 是在年龄 y_i 时面临身故风险的个体总数，每个个体风险事件发生的概率相互独立，因此风险事件发生的总数 s_i 服从二项分布：

$$s_i\sim B(r_i,1-P\{T>y_i|T>y_{i-1}\})$$

其中，$P\{T>y_i|T>y_{i-1}\}=\dfrac{S(y_i)}{S(y_{i-1})}$。因此，

$$E\left(\frac{r_i-s_i}{r_i}\right)=\frac{r_i-r_i[1-S(y_i)/S(y_{i-1})]}{r_i}=\frac{S(y_i)}{S(y_{i-1})} \qquad (8.3.8)$$

$$Var\left(\frac{r_i-s_i}{r_i}\right)=\frac{Var(s_i)}{r_i^2}=\frac{r_i[1-S(y_i)/S(y_{i-1})]S(y_i)/S(y_{i-1})}{r_i^2}$$

$$=\frac{[S(y_{i-1})-S(y_i)]S(y_i)}{r_jS(y_{i-1})^2} \qquad (8.3.9)$$

在指定的死亡时刻 y_j 处的生存概率估计值 $S_n(y_j)$ 是无偏的，这是因为：

$$E[S_n(y_j)] = E\left(\prod_{i=1}^{j}\frac{r_i - s_i}{r_i}\right) = \prod_{i=1}^{j}E\left(\frac{r_i - s_i}{r_i}\right) = \prod_{i=1}^{j}\frac{S(y_i)}{S(y_{i-1})} = \frac{S(y_j)}{S(y_0)}$$

而在没有死亡事件发生的时刻，生存概率 $S(t)$ 的估计值是有偏的，偏误为：

$$E[S_n(t)] - S(t) = S(y_{j-1}) - S(t), \qquad y_{j-1} \leqslant t < y_j$$

下面我们计算在死亡时刻 y_j 处估计量的方差。推导的过程需要用到如下两个性质：

1. 设 $X_1, X_2, \cdots, X_n$ 是相互独立的随机变量，$E(X_i) = \mu_i$，$Var(X_i) = \sigma_i^2$。则有

$$\begin{aligned} Var(X_1 \cdots X_n) &= E(X_1^2 X_2^2 \cdots X_n^2) - [E(X_1 X_2 \cdots X_n)]^2 \\ &= E(X_1^2)E(X_2^2)\cdots E(X_n^2) - (EX_1)^2(EX_2)^2\cdots(EX_n)^2 \\ &= \prod_{i=1}^{n}(\mu_i^2 + \sigma_i^2) - \prod_{i=1}^{n}\mu_i^2 \end{aligned}$$

2. 若 a_i，$i = 1, 2, \cdots, n$ 相对于某一常数 p 均为高阶无穷小，$a_i \sim o(p)$，则

$$\prod_{i=1}^{n}(1 + a_i) = 1 + \sum_{i=1}^{n}a_i + o(p^2)$$

由以上两条性质，即可进行方差的近似计算：

$$\begin{aligned} Var[S_n(y_j)] &= Var\left[\prod_{i=1}^{j}\left(\frac{r_i - s_i}{r_i}\right)\right] \\ &= \prod_{i=1}^{j}\left\{\frac{S(y_i)^2}{S(y_{i-1})^2} + \frac{[S(y_{i-1}) - S(y_i)]S(y_i)}{r_i S(y_{i-1})^2}\right\} - \prod_{i=1}^{j}\left[\frac{S(y_i)^2}{S(y_{i-1})^2}\right] \\ &= \prod_{i=1}^{j}\left\{\frac{r_i S(y_i)^2 + [S(y_{i-1}) - S(y_i)]S(y_i)}{r_i S(y_{i-1})^2}\right\} - \frac{S(y_j)^2}{S(y_0)^2} \\ &= \prod_{i=1}^{j}\left\{\frac{S(y_i)^2}{S(y_{i-1})^2}\frac{r_i S(y_i) + [S(y_{i-1}) - S(y_i)]}{r_i S(y_i)}\right\} - \frac{S(y_j)^2}{S(y_0)^2} \\ &= \frac{S(y_j)^2}{S(y_0)^2}\left\{\prod_{i=1}^{j}\left[1 + \frac{S(y_{i-1}) - S(y_i)}{r_i S(y_i)}\right] - 1\right\} \\ &\approx \frac{S(y_j)^2}{S(y_0)^2}\sum_{i=1}^{j}\frac{S(y_{i-1}) - S(y_i)}{r_i S(y_i)} \end{aligned} \qquad (8.3.10)$$

注意到 $\frac{r_i - s_i}{r_i}$ 是 $\frac{S(y_i)}{S(y_{i-1})}$ 的估计，将其带入式（8.3.10），就得到了 Greenwood 近似公式（Greenwood approximation）：

$$\widehat{Var}[S_n(y_j)] = S_n(y_j)^2 \sum_{i=1}^{j}\frac{s_i}{r_i(r_i - s_i)} \qquad (8.3.11)$$

Greenwood 近似公式可以拓展到一般的时刻 t 的情形：

$$\widehat{Var}[S_n(t)]=S_n(t)^2\sum_{i:y_i\leq t}\frac{s_i}{r_i(r_i-s_i)} \tag{8.3.12}$$

【例 8－18】 利用例 8－2 中的数据，用 Greenwood 近似公式估计${}_2p$ 和 q_2 的方差。

解： 对于例 8－2，$r_1=30$，$s_1=1$，$r_2=29$，$s_2=1$。

$$\widehat{Var}[S_{30}(2)]=\left(\frac{28}{30}\right)^2\left(\frac{1}{30\cdot 29}+\frac{1}{29\cdot 28}\right)=\frac{56}{30^3}$$

若使用经验估计公式（8.2.17），则有

$$\widehat{Var}[S_{30}(2)]=\frac{(2/30)(28/30)}{30}=\frac{56}{30^3}$$

二者的估计结果是相同的。

对于${}_{2|}q$ 的估计，注意到方差的计算必须在给定第二年初有 28 人的情况下进行。从此刻开始，只有一次死亡事件发生：$r_1=28$，$s_1=1$。因此，

$$\widehat{Var}\left({}_{2|}\hat{q}\mid {}_2p=\frac{28}{30}\right)=\left(\frac{27}{28}\right)^2\frac{1}{28\cdot 27}=\frac{27}{28^3}$$

注意到这个结果和例 8－2 的结果是一致的。■

从本例可以看出，Greenwood 公式是完整个体数据的方差估计的推广。事实上，在完整样本的情形，$r_{i+1}=r_i-s_i$，当 $y_j\leq t<y_{j+1}$时，

$$\begin{aligned}\widehat{Var}[S_n(t)]&=S_n(t)^2\sum_{i:y_i\leq t}\frac{s_i}{r_i(r_i-s_i)}=S_n(t)^2\sum_{i:y_i\leq t}\frac{r_i-r_{i+1}}{r_ir_{i+1}}\\&=S_n(t)^2\sum_{i:y_i\leq t}\left(\frac{1}{r_{i+1}}-\frac{1}{r_i}\right)=\left(\frac{r_j}{n}\right)^2\left(\frac{1}{r_j}-\frac{1}{n}\right)\\&=\frac{r_j(n-r_j)}{n^3}=\frac{1}{n}S_n(t)(1-S_n(t))\end{aligned}$$

结果和例 8－10 完全相同。

【例 8－19】 考虑例 8－5 的数据，重新估计 $S_n(2)$ 的方差，并且构造 95％的置信区间。

解： 从例 8－16 的结果可以知道：$r_1=r_2=29$，$s_1=s_2=1$。由例 8－17 知道，$S_{40}(2)=0.9322$。代入 Greenwood 公式，有

$$\widehat{Var}[S_{40}(2)]=(0.9322)^2\left(\frac{1}{29\cdot 28}+\frac{1}{29\cdot 28}\right)=0.002140$$

运用正态近似，得到的置信区间端点为：$0.9322\pm 1.96\sqrt{0.002140}$，得到的区间为（0.8415，1.0229）。■

注意到，在例 8－19 中，置信区间的上界为 1.0229 大于 1，这个值是没有意义的。在小样本情形下，直接运用正态近似，极有可能得到这样的结果。解决这个问题的方法是使用**对数转换的置信区间**（log-transformed confidence interval）进行估计。

对数转换的基本原理如下：令 $Y_n = g(X_n)$，其中 g（·）是连续可微的函数，X_n 是对 $E(X_n)$ 的一致估计。则在适当的正则性条件下，

$$Y_n - E(Y_n) = g(X_n) - E(g(X_n)) = g(X_n) - g(E(X_n)) + g(E(X_n)) - E(g(X_n))$$
$$= g'(X_n)(X_n - E(X_n)) + o(X_n - E(X_n))$$

因此，$Var(Y_n) \approx [g'(X_n)]^2 Var(X_n)$。

我们取 $Y = \ln[-\ln(S_n(t))]$，相应地 $g'(x) = \frac{1}{x\ln x}$。而 Kaplan - Meier 估计 $S_n(t)$ 是 $S(t)$ 的一致估计。因此，$\widehat{Var}(Y) = \frac{\widehat{Var}[S_n(t)]}{[S_n(t)\ln S_n(t)]^2}$。由此估计 $\theta = \ln[-\ln S(t)]$ 的端点为：

$$\ln[-\ln S_n(t)] \pm 1.96\frac{\sqrt{\widehat{Var}[S_n(t)]}}{S_n(t)\ln S_n(t)}$$

对应的 $S(t)$ 的置信区间端点为：

$$\exp\left\{-\exp\left(\ln[-\ln S_n(t)] \pm 1.96\frac{\sqrt{\widehat{Var}[S_n(t)]}}{S_n(t)\ln S_n(t)}\right)\right\} = S_n(t)^{\exp\left\{\pm 1.96\frac{\sqrt{\widehat{Var}[S_n(t)]}}{S_n(t)\ln S_n(t)}\right\}}$$

这样得到的区间总在（0，1）之内。

【例 8 - 20】 用对数转换方法重新计算例 8 - 19。

解： 为表示方便，我们记

$$U = \exp\left\{1.96\frac{\sqrt{\widehat{Var}[S_n(t)]}}{S_n(t)\ln S_n(t)}\right\} = \exp\left\{\frac{1.96 \cdot \sqrt{0.002140}}{0.9322 \cdot \ln 0.9322}\right\} = 0.2502$$

所以置信区间为 $(S_n(t)^{U^{-1}}, S_n(t)^{U}) = (0.9322^{0.2502^{-1}}, 0.9322^{0.2502}) = (0.7554, 0.9826)$。■

8.3.3 危险率函数的 Nelson-Åalen 估计

前文介绍的经验模型是离散分布模型，根据这个离散模型无法求导获得密度函数和危险率函数。下面介绍修正的 Nelson-Åalen 方法来估计累积危险率函数 $H(t)$。在第二章中，我们已经知道 $H(t)$、$h(t)$ 和 $S(t)$ 满足如下关系式：

$$H(t) = -\ln S(t) \tag{8.3.13}$$

$$H'(t) = -\frac{S'(t)}{S(t)} = \frac{f(t)}{S(t)} = h(t) \tag{8.3.14}$$

下面给出 $H(t)$ 估计的直观推导。在任意时刻 t，令 $r(t)$ 是风险集，$h(t)$ 是危险率函数，$s(t)$ 表示在时刻 t 之前身故的个体数目的期望值，则可以得出以下结论：

$$s(t) = \int_0^t r(u)h(u)\mathrm{d}u$$

两边求导，有 $ds(t)=r(t)h(t)dt$。于是，$\frac{ds(t)}{r(t)}=h(t)dt$。两边积分，得：

$$\int_0^t \frac{ds(t)}{r(t)} = \int_0^t h(t)dt = H(t)$$

现在将真实的期望值 $s(t)$ 用 t 时刻之前的身故个数的观测值$\hat{s}(t)$代替。$\hat{s}(t)$ 是一个阶梯函数，在每一个身故发生的时刻增加 s_i，因此上式的左边变为 $\sum_{t_i\leqslant t}\frac{s_i}{r_i}$。这就给出了 $H(t)$ 的 Nelson-Åalen 估计值：

$$\hat{H}(x)=\sum_{j:y_j\leqslant x}\frac{s_j}{r_j}=\begin{cases}0, & x<y_1\\ \sum_{j=1}^{i-1}\frac{s_j}{r_j}, & y_{i-1}\leqslant x<y_i\\ \sum_{j=1}^{k}\frac{s_j}{r_j}, & x\geqslant y_k\end{cases} \tag{8.3.15}$$

相应地，由式（8.3.13）生存函数的 Nelson-Åalen 估计值为：

$$\hat{S}(x)=e^{-\hat{H}(x)}$$

为了计算 Nelson-Åalen 估计的方差，我们假设在身故时刻 y_j 附近的个体数服从参数 $r_jh(y_j)$的泊松分布。因此，$Var(s_j)=r_jh(t_j)$可以用 $r_j\frac{s_j}{r_j}=s_j$ 来估计。假设各时间点死亡人数相互独立，于是可近似地有

$$\widehat{Var}[\hat{H}(y_j)]=\widehat{Var}\left(\sum_{i=1}^{j}\frac{s_i}{r_i}\right)=\sum_{i=1}^{j}\frac{\widehat{Var}(s_i)}{r_i^2}=\sum_{i=1}^{j}\frac{s_i}{r_i^2} \tag{8.3.16}$$

对一般的时间 t，我们有：

$$\widehat{Var}[\hat{H}(t)]=\widehat{Var}\left(\sum_{y_i:y_i\leqslant t}\frac{s_i}{r_i}\right)=\sum_{y_i:y_i\leqslant t}\frac{s_i}{r_i^2} \tag{8.3.17}$$

由此，得到 $H(t)$的线性置信区间为：

$$\hat{H}(t)\pm\mu_{\alpha/2}\sqrt{\sum_{y_i:y_i\leqslant t}\frac{s_i}{r_i^2}}$$

和 Kaplan – Meier 的估计类似，可以推导出其对数转换的置信区间为：

$$\hat{H}(t)\exp\left\{\pm\frac{\mu_{\alpha/2}\sqrt{\sum_{y_i:y_i\leqslant t}\frac{s_i}{r_i^2}}}{\hat{H}(t)}\right\}$$

注意，这里我们没有假定数据的非完整性。无论对于完整数据还是非完整数据，Nelson-Åalen 估计都是相同的。

【例 8 – 21】 利用例 8 – 7 的数据，计算累积死亡力函数的 Nelson-Åalen 估计，与例 8 – 9 的结果作比较。

解： 由（8.3.15）计算得：

$$\hat{H}(x)=\begin{cases}0, & x<1\\ \frac{1}{8}=0.125, & 1\leqslant x<2\\ 0.125+\frac{2}{7}=0.4107, & 2\leqslant x<4\\ 0.4107+\frac{2}{5}=0.8107, & 4\leqslant x<6\\ 0.8107+\frac{1}{3}=1.1440, & 6\leqslant x<7\\ 1.1440+\frac{1}{2}=1.6440, & 7\leqslant x<9\\ 1.6440+\frac{1}{1}=2.6440, & x\geqslant 9\end{cases}$$

利用 Nelson-Åalen 估计得到的分布函数的估计$\hat{F}(x)=1-e^{-\hat{H}(x)}$与经验分布函数 $F_n(x)$ 是不同的，可在图 8-5 中看出。■

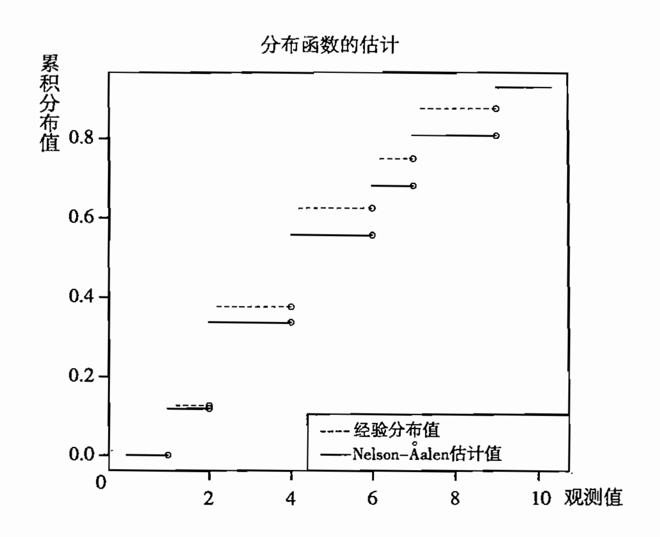


图 8-5　经验分布函数与 Nelson-Åalen 估计的比较

【例 8-22】　对例 8-5 中的数据，估计 $H(2)$95%的近似置信区间。

解： 由例 8-16 的结果，$\hat{H}(2)=1/29+1/29=0.06897$，方差估计值为 $1/29^2+1/29^2=0.002378$。线性置信区间的端点是：$0.06897\pm1.96\sqrt{0.002378}=0.06897\pm0.09558$，得到的区间是：$(-0.02661, 0.1645)$。注意到区间包括了负值。采用对数转换的方法，令

$$U = \exp\left\{\frac{z_{\frac{\alpha}{2}}\sqrt{\sum_{i=1}^{j} s_i / r_i^2}}{\hat{H}(y_j)}\right\} = \exp\left\{\frac{1.96\sqrt{0.002378}}{0.06897}\right\} = 3.9980$$

从而置信区间的端点为 $0.06897U^{\pm 1}$，得到的区间是（0.01725，0.27576）。

■

§8.4 核密度估计

经验分布估计作为对完整数据分布函数的估计，具有无偏性和一致性的良好性质。但是经验分布是一种离散分布，在每个观测点有点概率，其余点的概率为0。如果已知真实的分布是连续的，则经验分布的近似程度较差。**核密度估计方法**（kernel density estimation）用一种连续的分布函数来近似离散的经验分布函数，从而为分布密度的估计提供了一种手段。

8.4.1 核函数与核密度估计

我们沿用 8.2.1 中的记号。假设完整的数据集为$\{x_1, x_2, \cdots, x_n\}$，其中包括不同的数值为 $y_1 < y_2 < \cdots < y_k$，并且记 $s_j = \#\{x_i : x_i = y_j\}$ 表示取值 y_j 的观测值的个数，则观测值 y_j 的经验密度是 $p(y_j) = \frac{s_j}{n}$。由 8.2 中的结论知道，经验分布函数的估计为：

$$F_n(x) = \sum_{i=1}^{n} \frac{1}{n} I(x_i \leqslant x) = \sum_{j: y_j \leqslant x} p(y_j)$$

核密度估计法使用连续随机变量来替代每个离散的点，使用核函数 $k(\cdot)$进行核密度估计的估计量是：

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k\left(\frac{x_i - x}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} k_h(x_i - x) \qquad (8.4.1)$$

其中 h 是带宽（bandwidth），$k_h(\cdot) = \frac{1}{h} k\left(\frac{\cdot}{h}\right)$是尺度变换后的核函数。

如果用 $y_1 < y_2 < \cdots < y_k$ 来表示，式（8.4.1）也可写做：

$$\hat{f}_h(x) = \sum_{j=1}^{k} p(y_j) k_h(y_j - x) = \int_{-\infty}^{\infty} k_h(y - x) dF_n(y) \qquad (8.4.2)$$

显然两种表示方法是等价的。

使得$\hat{f}_h(x)$ 成为密度函数的充要条件是用于加权的核函数 $k_h(\cdot)$ 是一个密度函数。这是因为：

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \hat{f}_h(x) dx = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} k_h(y - x) dF_n(y) = \int_{-\infty}^{\infty} dF_n(y) \int_{-\infty}^{\infty} k_h(y - x) dx \\ &= \int_{-\infty}^{\infty} dF_n(y) \int_{-\infty}^{\infty} k_h(u) d(u - x) = \int_{-\infty}^{\infty} dF_n(y) \int_{-\infty}^{\infty} k_h(u) du = \int_{-\infty}^{\infty} k_h(u) du \end{aligned}$$

习惯上要求密度函数为核函数的随机变量的均值等于所替代的数据点，但这并不是必须的，而仅仅是为了让核估计与经验估计有相同的均值。

进行核密度估计时，核函数的形式 $k(\cdot)$ 和带宽 h 都是事先给定的。为了研究的方便，我们假设 $k(\cdot)$ 仅在闭区间［-1，1］上取非零值（具有紧支集），而尺度参数 h 使得 $k_h(\cdot)$ 的形状可以拓展到一般的形式。另一方面，把 h 明确地写成一个核函数的参数，可以独立地分析带宽对估计方差和偏误的影响。

由密度函数的核密度估计，可以得到分布函数的核密度估计。当 $k(\cdot)$ 是对称分布时，

$$\hat{F}_h(x)=\int_{-\infty}^{x}\hat{f}_h(z)dz=\int_{-\infty}^{x}dz\int_{-\infty}^{\infty}k_h(y-z)dF_n(y)=\int_{-\infty}^{\infty}dF_n(y)\int_{-\infty}^{x}k_h(y-z)dz$$

$$=\int_{-\infty}^{\infty}K_h(x-y)dF_n(y)=\sum_{j:y_j\leqslant x}p(y_j)K_h(x-y_j) \tag{8.4.3}$$

其中，$K_h(\cdot)$ 是密度函数 $k_h(\cdot)$ 对应的分布函数：$K_h(x)=\int_{-\infty}^{x}k_h(u)du$。如果令 $K(x)=\int_{-\infty}^{x}k(u)du$，则易知 $K_h(x)=K\left(\frac{x}{h}\right)$。

8.4.2　常见的核函数

从式（8.4.2）可以知道，任意的密度函数都可以作为核函数进行核密度估计。常见的核函数有以下三种：

1. 均匀核函数。如果我们取［-1，1］上均匀分布的密度函数作为核函数，则有

$$k(u)=\frac{1}{2},\quad -1\leqslant u\leqslant 1$$

加入带宽作为尺度参数后，一般的均匀核函数为：

$$k_h(u)=\frac{1}{2h},\quad -h\leqslant u\leqslant h \tag{8.4.4}$$

它对应了［-h，h］上的均匀分布。相应地，

$$K_h(u)=\begin{cases}0, & u<-h\\ \dfrac{u+h}{2h}, & -h\leqslant u\leqslant h\\ 1, & u>h\end{cases} \tag{8.4.5}$$

基于均匀核函数的密度函数的核密度估计为：

$$\hat{f}_h(x)=\sum_{i=1}^{n}\frac{1}{2nh}I(x_i-h\leqslant x\leqslant x_i+h)=\sum_{j=1}^{k}p(y_j)\frac{1}{2h}I(y_j-h\leqslant x\leqslant y_j+h) \tag{8.4.6}$$

分布函数的核密度估计为：

$$\hat{F}_h(x) = \sum_{j:x-h\leqslant y_j\leqslant x+h} p(y_j)\frac{x-y_j+h}{2h} + \sum_{j:y_j<x-h} p(y_j) \tag{8.4.7}$$

在均匀核函数下，密度函数的估计就是对每一点带宽范围内的经验密度的平均值，它和直方图既有联系又有区别。把密度函数的估计可以写成：

$$\hat{f}_h(x) = \frac{1}{2nh}\#\{x_i: x-h\leqslant x_i\leqslant x+h\}$$

而直方图的计算公式可以写做：

$$f_n(x) = \frac{\#\{x_i: c_{j-1}\leqslant x_i < c_j\}}{n(c_j - c_{j-1})}, \qquad c_{j-1}\leqslant x < c_j$$

容易看出，二者本质都是对一定范围内的经验密度求取平均，估计的密度函数都是分段常值的函数。两种方法的区别是，在直方图中每个分段区间 $[c_{j-1}, c_j)$ 是事先给定的，与 x 无关，而在核密度估计中分段区间 $[x-h, x+h]$ 是由 x 决定的。

【例 8－23】 某个损失数据的样本为：7，12，15，19，26，27，29，29，30，33，38，53。给定带宽参数 $h=5$，利用均匀核函数，估计 $\hat{f}(20)$，$\hat{F}(20)$，$\hat{f}(30)$，$\hat{F}(30)$。

解： 我们借助图 8－6 说明核密度估计的原理。左图是密度函数的估计，右图是分布函数的估计，虚线划分的范围表示估计点 20 和 30 所对应的带宽。对于密度估计，只需要将带宽范围内的点利用对应的核函数加权平均；对于分布函数估计，除了带宽范围内的点加权以外，还要对左侧所有点赋予 1 的权重。

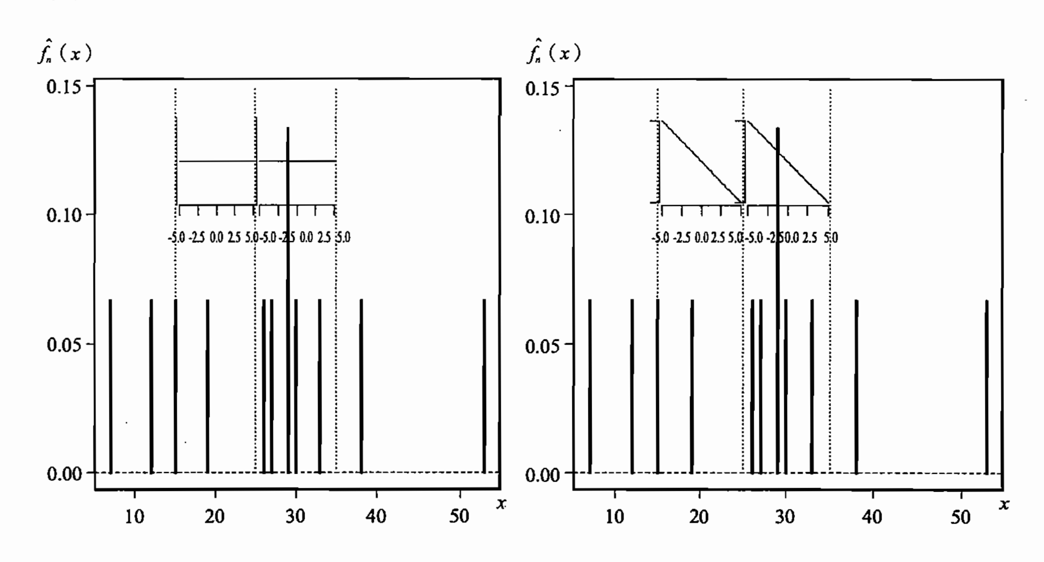


图 8－6 均匀核密度估计示意图

对于$\hat{f}(20)$，需要考虑15，19处的经验密度，两点的权重都为$\frac{1}{2h}=\frac{1}{10}$。利用式（8.4.6）得到：

$$\hat{f}(20)=\frac{1}{12}\cdot\frac{1}{10}+\frac{1}{12}\cdot\frac{1}{10}=\frac{1}{60}$$

对于$\hat{F}(20)$，需要考虑的点包括7、12、15、19。其中前两个点在带宽范围左边，权重为1，利用式（8.4.7），第三个点权重也为1，第四个点权重为0.6。因此：

$$\hat{F}(20)=\frac{1}{12}\cdot 1+\frac{1}{12}\cdot 1+\frac{1}{12}\cdot 1+\frac{1}{12}\cdot 0.6=\frac{3}{10}$$

同法可得：

$$\hat{f}(30)=\frac{1}{12}\cdot\frac{1}{10}+\frac{1}{12}\cdot\frac{1}{10}+\frac{1}{12}\cdot\frac{1}{10}+\frac{1}{12}\cdot\frac{1}{10}+\frac{1}{12}\cdot\frac{1}{10}+\frac{1}{12}\cdot\frac{1}{10}=\frac{1}{20}$$

$$\hat{F}(30)=\frac{1}{12}\cdot 1+\frac{1}{12}\cdot 1+\frac{1}{12}\cdot 1+\frac{1}{12}\cdot 1+\frac{1}{12}\cdot 0.9+\frac{1}{12}\cdot 0.8+\frac{1}{12}\cdot 0.6+\frac{1}{12}\cdot 0.5+\frac{2}{12}\cdot 0.2=\frac{19}{30}$$

我们可以绘制核密度估计的分布函数和密度函数图像，如图8－7所示。

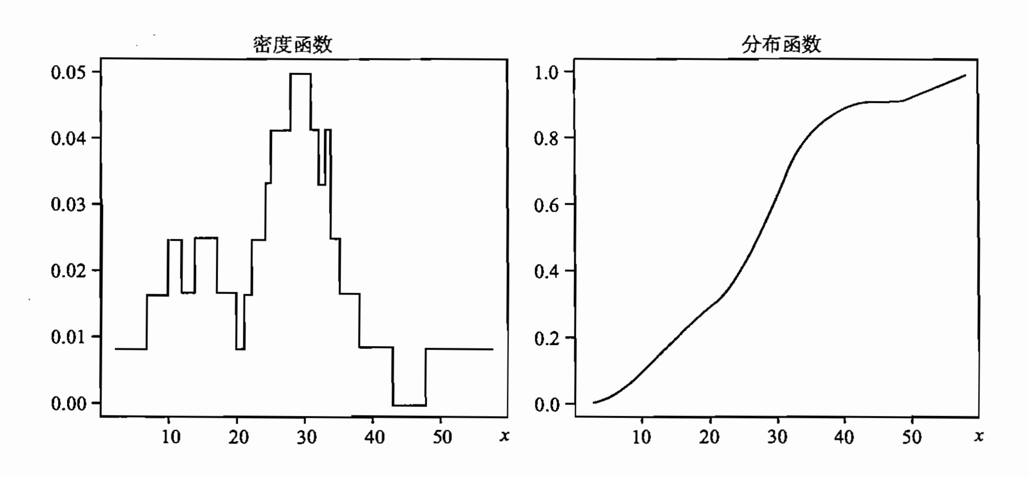


图8－7 均匀核密度估计的密度函数和分布函数

■

2. 三角核函数。如果我们取［－1，1］上三角形的函数作为核函数，则有

$$k(u)=1-|x|,\quad -1\leqslant x\leqslant 1$$

$$K(u)=\begin{cases}0, & u<-1\\ \dfrac{(u+1)^2}{2}, & -1\leqslant u<0\\ 1-\dfrac{(u-1)^2}{2}, & 0\leqslant u<1\\ 1, & u\geqslant 1\end{cases}$$

加入带宽作为尺度参数后，一般的三角核函数为：

$$k_h(u)=\frac{h-|u|}{h^2},\quad -h\leqslant u\leqslant h \tag{8.4.8}$$

它对应了 $[-h,\ h]$ 上的三角形分布。相应地，

$$K_h(u)=K\left(\frac{u}{h}\right)=\begin{cases}0, & u<-h\\ \dfrac{(u+h)^2}{2h^2}, & -h\leqslant u<0\\ 1-\dfrac{(u-h)^2}{2h^2}, & 0\leqslant u<h\\ 1, & u\geqslant h\end{cases} \tag{8.4.9}$$

【例 8－24】 重新计算例 8－23，利用三角核函数估计，带宽不变。

解： 这里也给出相应的估计示意图，见图 8－8。

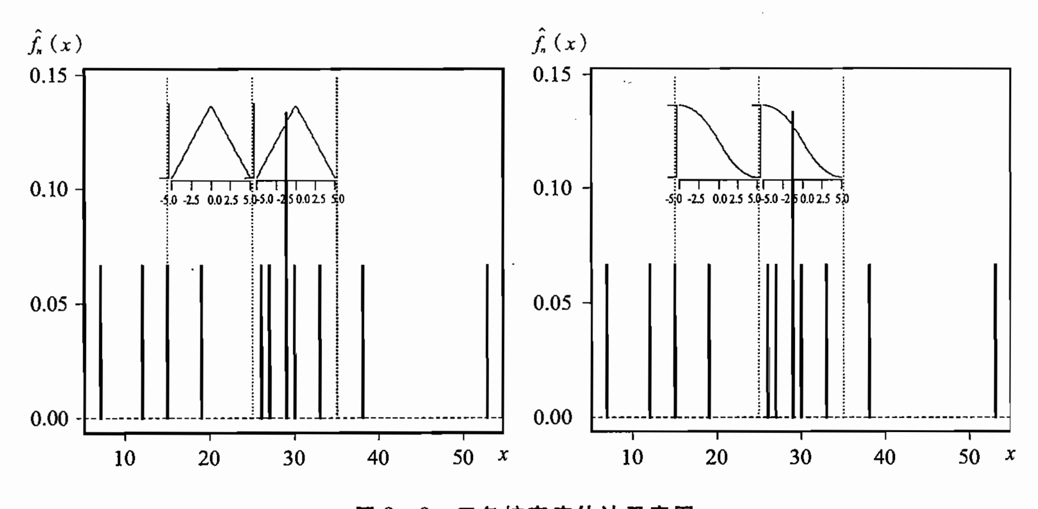


图 8－8 三角核密度估计示意图

估计过程类似 8－23。为计算 $\hat{f}(20)$，需要考虑的点是 15，19。由式 (8.4.8) 得到相应的权重分别为 0 和 $\dfrac{5-|20-19|}{5^2}=\dfrac{4}{25}$。因此：

$$\hat{f}(20)=\frac{1}{12}\cdot\frac{4}{25}=\frac{1}{75}$$

对于对于 $\hat{F}(20)$，由式 (8.4.9) 计算得到点 7、12、15、19 的权重分别为

1、1、1、$1-\frac{(19-20+5)^2}{2\cdot 5^2}=\frac{17}{25}$。因此，

$$\hat{F}(20)=\frac{1}{12}\cdot 1+\frac{1}{12}\cdot 1+\frac{1}{12}\cdot 1+\frac{1}{12}\cdot\frac{17}{25}=\frac{23}{75}$$

同理可得，$\hat{f}(30)=3/50$，$\hat{F}(30)=49/75$。相应的估计结果示意图如图 8－9 所示。

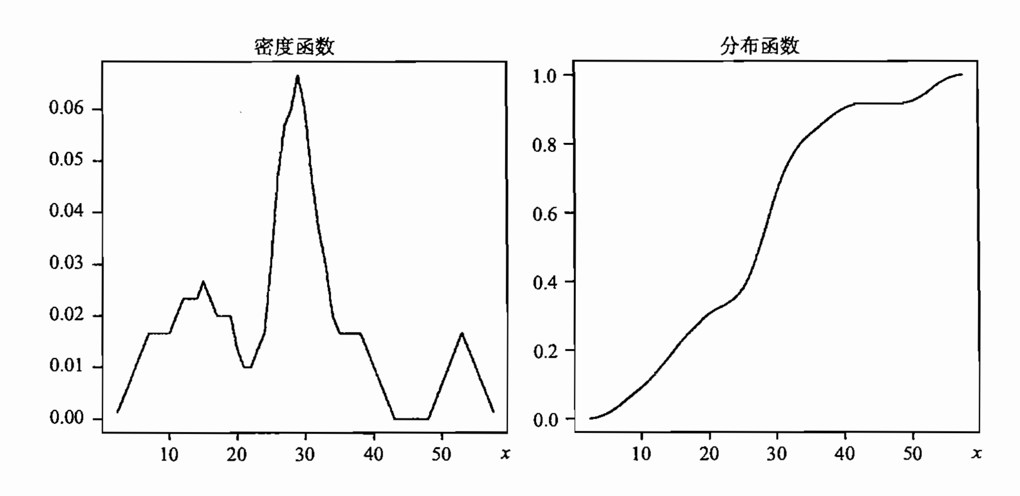


图 8－9　三角核密度估计的密度函数和分布函数

■

3. *伽玛核函数*。前文介绍的核函数是一般的情形，在给定带宽的情况下，权重仅和观测值与待估计点 x 的距离有关。伽玛核函数是一种特殊的核函数，它的权重不是按照样本点与待估计点的距离来定义的。对于每一个观测值 x_i，我们用一个均值为 x_i 的伽玛分布 Γ（α，x_i/α）来代表，其密度函数为：

$$k_{x_i}(x)=\frac{x^{\alpha-1}e^{-x\alpha/x_i}}{(x_i/\alpha)^{\alpha}\Gamma(\alpha)} \tag{8.4.10}$$

任意待估计点 x 的密度可以通过这样一族伽玛分布的平均值得到：

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{x^{\alpha-1}e^{-x\alpha/x_i}}{(x_i/\alpha)^{\alpha}\Gamma(\alpha)}=\sum_{j=1}^{k}p(y_j)\frac{x^{\alpha-1}e^{-x\alpha/y_j}}{(y_j/\alpha)^{\alpha}\Gamma(\alpha)} \tag{8.4.11}$$

容易知道，在这个核密度估计中，参数 α 是尺度参数，α 越大对应的估计越平滑。而最终估计的函数是一个混合的伽玛分布。

【例 8－25】　重新计算例 8－23 中的两个密度估计，利用伽玛核函数，α 取 5 或 50。

解： 伽玛核函数估计中，对任意的待求点，所有的观测值都将赋予权重。因此计算量相对较大，此处省略中间步骤。当 $\alpha=5$ 时，$\hat{f}(20)=0.3200$，$\hat{f}(30)=0.2366$；当 $\alpha=50$ 时，$\hat{f}(20)=0.2111$，$\hat{f}(30)=0.5029$。

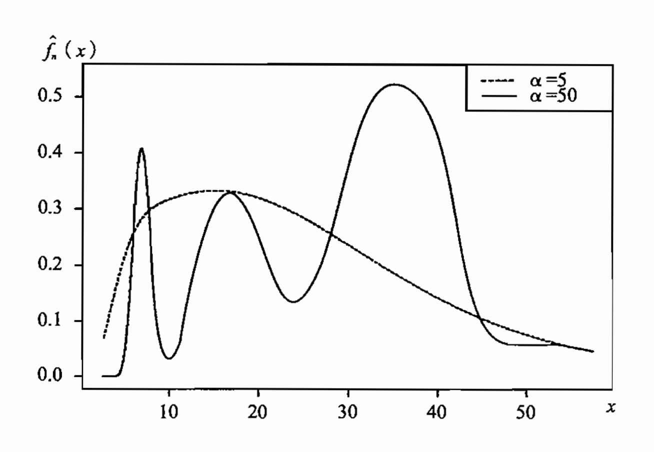


图 8－10　不同尺度参数下伽玛核密度估计

图 8－10 列出了两种尺度参数下的估计结果。从图中可以看出，在不同尺度参数下，估计的形态是显著不同的。■

8.4.3　带宽对估计的影响

从例 8－25 中，我们已经看到尺度参数对估计的影响。在一般核函数估计中，实践中证明核函数的选取对估计的偏误影响不大，而带宽 h 的变化却会使估计结果产生很大的波动。

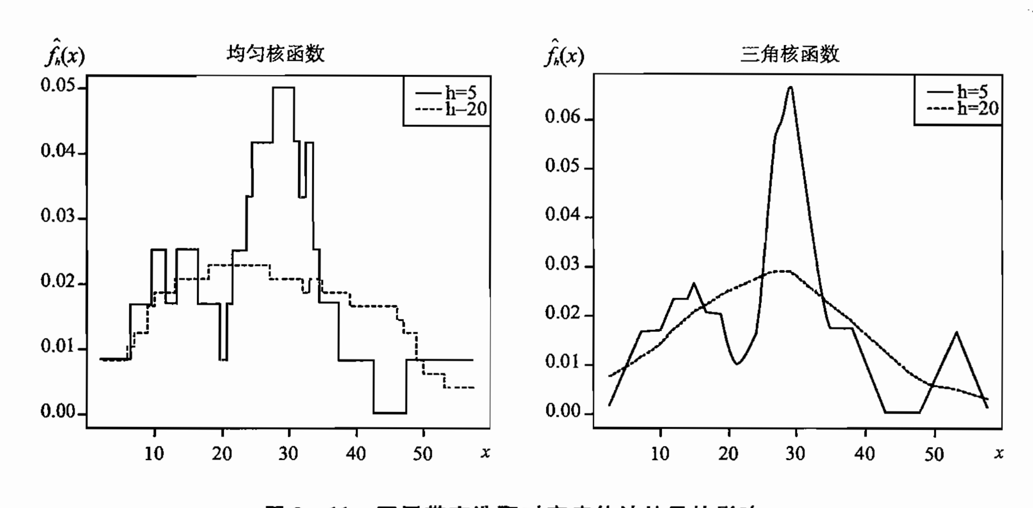


图 8－11　不同带宽选取对密度估计结果的影响

图 8－11 绘制了分别取带宽 $h=5$，$h=20$ 对例 8－23 的数据进行估计的结果。可以看出，大的带宽可以得到更加光滑的结果，而小的带宽能够反映出局部的密度变化。对于相同的带宽，采用不同核函数估计的结果相差并不大。理论上，带宽造成的渐进偏差为 $O(h^2)$，而渐进方差为 $O\left(\frac{1}{nh}\right)$，

因此使估计偏误最小的带宽为 $O(n^{-\frac{1}{5}})$，所以样本量增大，相应应该选取较小的带宽。实际应用中，需要根据一定的规则进行选取。

§8.5　大样本数据下的经验分布函数估计

当数据量很大的时候，如果使用 Kaplan - Meier 估计方法，必须进行大量的数据排序和计算工作，但是从结果看，如此繁杂的工作是不必要的。比如在生命表的估计中，我们只需要计算整数年龄的函数值，而不需要知道每个观测值的生存函数。因此，可以采用近似的方法进行估计。

8.5.1　Kaplan - Meier 近似

给定区间的端点 $c_0 < c_1 < \cdots < c_k$，令 $D_j = \#\{d_i : c_{j-1} \leqslant d_i < c_j\}$ 是区间 $[c_{j-1}, c_j)$ 中某个点的左截断的观测值个数，$U_j = \#\{u_i : c_{j-1} < u_i \leqslant c_j\}$ 表示区间 $(c_{j-1}, c_j]$ 上某个点右删失的观测值的个数。注意这里区间的处理是不同的，这是因为：如果截断发生在 c_j，或者删失发生在 c_{j-1}，则对于这个个体我们对它在区间 (c_{j-1}, c_j) 上的生存状况一无所知，因此对相应的函数估计没有影响。相应地，我们记 $(c_{j-1}, c_j]$ 中未删失的观测值数目为 X_j。此时，样本总量为 $n = \sum_{i=1}^{k} D_j = \sum_{j=1}^{k} (X_j + U_j)$。

风险集的计算理论上应该使用式（8.3.2），但是这里我们要求的是一个时间段上的风险集的大小。近似计算的思想是引入一个假定的值 $c_j^* \in (c_{j-1}, c_j)$，并且假设所有未删失事件都发生在 c_j^* 处。这样，我们就可以用 c_j^* 处的风险集来近似 (c_{j-1}, c_j) 上的风险集。则式（8.3.2）可以化为：

$$
\begin{aligned}
r_j &= \#\{d_i : d_i < c_j^*\} - \#\{x_i : x_i < c_j^*\} - \#\{u_i : u_i < c_j^*\} \\
&= [\#\{d_i : d_i < c_{j-1}\} - \#\{x_i : x_i < c_{j-1}\} - \#\{u_i : u_i < c_{j-1}\}] \\
&\quad + \#\{d_i : c_{j-1} \leqslant d_i < c_j^*\} - \#\{u_i : c_{j-1} \leqslant u_i < c_j^*\} \\
&= P_j + \#\{d_i : c_{j-1} \leqslant d_i < c_j^*\} - \#\{u_i : c_{j-1} \leqslant u_i < c_j^*\}
\end{aligned} \tag{8.5.1}
$$

其中，
$$P_j = \sum_{i=1}^{j-1} (D_i - X_i - U_i) \tag{8.5.2}$$

注意到式（8.5.1）中，在 c_{j-1}时刻风险集的大小 P_j 是可以计算的，而其他三项需要基于一定的假设进行调整。一般情况下，我们假设在 $[c_{j-1}, c_j^*)$ 上截断的个体数占 $[c_{j-1}, c_j)$ 上截断个体数 D_j 的比例是 $100\alpha\%$，在 $[c_{j-1}, c_j^*)$ 上删失的个体数占 $(c_{j-1}, c_j]$ 上删失个体数 U_j 的比例是 $100\beta\%$。这样，就有：

$$\#\{d_i : c_{j-1} \leqslant d_i < c_j^*\} = D_j \cdot 100\alpha\%$$

$$\#\{u_i : c_{j-1} \leqslant u_i < c_j^*\} = U_j \cdot 100\beta\%$$

代入式（8.5.1），即得到：

$$r_j = P_j + \alpha D_j - \beta U_j \tag{8.5.3}$$

Kaplan - Meier 估计中，每一项条件概率

$$\hat{P}(X > c_i \mid X > c_{i-1}) = 1 - \frac{X_i}{r_i}$$

因此，生存函数的估计为：

$$\hat{S}(c_j) = \prod_{i=1}^{j} P(X > c_i \mid X > c_{i-1}) = \prod_{i=1}^{j}\left(1 - \frac{X_i}{r_i}\right) \tag{8.5.4}$$

常见的一种选择是 $\alpha = 1$，$\beta = 0$，表示假设所有的左截断发生在 c_j^* 以前，同时所有的删失发生在 c_j^* 以后。此时，$r_i = \sum_{i=1}^{j} D_i - \sum_{i=1}^{j-1}(X_i + U_i)$。传统的生命表估计就是这种情形，此时分组 $c_0 < c_1 < \cdots < c_k$ 取一组整数值。死亡率的估计公式是：

$$q_j' \triangleq {}_{c_j - c_{j-1}}\hat{q}_{c_{j-1}} = \frac{\hat{S}(c_{j-1}) - \hat{S}(c_j)}{\hat{S}(c_{j-1})} = \frac{X_j}{r_j} \tag{8.5.5}$$

在此公式中，所有当前时段或者更早时候进入研究项目的个体都被视做可能身故的个体，而所有当前时段前就离开的个体不考虑在内。

另一种选择是令 $\alpha = \beta = 0.5$，这相当于假设进入和退出研究项目的事件在区间内均匀分布，而所有风险事件发生在区间的中点。由于 $P_{j+1} = P_j + D_j - U_j - X_j$，因此式（8.5.3）也可写做：

$$r_j = 0.5(P_j + P_{j+1} + X_j) \tag{8.5.6}$$

8.5.2 多元衰减表

在第二章中，我们考虑多因素死亡模型，分别计算各个因素造成的死亡概率。这个概念可以推广到非寿险模型中，在一次研究项目中，可以同时考虑多个因素同时造成的观测值数目的衰减，计算**多元终止概率**（multiple - decrement probability）。我们用 $q_j'^{(i)}$ 表示单个因素 i 引起的终止概率。例如，在生命表估计中，衰减可以由身故 d、退保 w 和退休 r 造成的，则可以用 $q'^{(w)}_j$ 表示在身故和退休都没有发生的前提下，年龄为 c_j 的个体在 c_{j+1} 前退保的概率，而 $q_j^{(w)}$ 表示在身故和退休可能发生的情况下相应的退保概率。

在第二章中，我们已经讨论了在不同假设下，$q_j'^{(i)}$ 和 $q_j^{(i)}$ 的相互推算关系。此处我们仅考虑在均匀分布假设下终止概率的估计，这时，

$$q_j^{(i)} = \frac{\ln(1 - q_j'^{(i)})}{\ln(1 - q_j^{(\tau)})} q_j^{(\tau)} \tag{8.5.7}$$

其中 $q_j^{(\tau)} = 1 - \prod_{i=1}^{m}(1 - q_j'^{(i)})$ 表示总终止概率。

【例 8－26】　运用例 8－5 中的数据，以整数年份划分区间，根据死亡时间均匀分布假设，估计相应的单个衰减因子和多元衰减相关量。

解： 注意对端点的处理。在 0 处截断的数据实际上是没有截断的，因此 $D_0=30$，而 $D_1=1$。根据死亡时间均匀分布的假设，就可以得到 $r_1=P_1+0.5(D_1-U_1)=29.5$。其他区间计算利用式（8.5.2）、（8.5.5）和（8.5.6）计算。结果如表 8－9 所示。

表 8－9　　例 8－26 的 Kaplan－Meier 近似计算法估计生存函数

c_j	D_j	U_j	X_j	P_j	r_j	$q_j'^{(d)}(=X_j/r_j)$	$\hat{S}(c_j)$
0	30	0	0	—	—	0.0000	1.0000
1	1	2	1	30	29.5	0.0339	0.9661
2	1	1	1	28	28.0	0.0357	0.9316
3	1	0	1	27	27.5	0.0364	0.8977
4	1	2	0	27	26.5	0.0000	0.8977
5	1	2	0	26	25.5	0.0000	0.8977
6	1	1	1	25	25.0	0.0400	0.8618
7	2	2	2	24	24.0	0.0833	0.7900
8	2	2	2	22	22.0	0.0909	0.7182
9	0	1	2	20	19.5	0.1026	0.6445
10	0	17	0	17	8.5	0.0000	0.6445

对于退保的情形，可以依据式（8.5.7）进行计算，如表 8－10 所示。

表 8－10　　例 8－26 的多元衰减估计

c_j	U_j	X_j	r_j	$q_j'^{(d)}$	$q_j'^{(w)}(=U_j/r_j)$	$q_j^{(\tau)}$	$q_j^{(d)}$	$q_j^{(w)}$
0	0	0	—	0.0000	0.0000	0.0000	—	—
1	2	1	29.5	0.0339	0.0678	0.0994	0.0327	0.0667
2	1	1	28.0	0.0357	0.0357	0.0702	0.0351	0.0351
3	0	1	27.5	0.0364	0.0000	0.0364	0.0364	0.0000
4	2	0	26.5	0.0000	0.0755	0.0755	0.0000	0.0755
5	2	0	25.5	0.0000	0.0784	0.0784	0.0000	0.0784
6	1	1	25.0	0.0400	0.0400	0.0784	0.0392	0.0392
7	2	2	24.0	0.0833	0.0833	0.1597	0.0799	0.0799
8	2	2	22.0	0.0909	0.0909	0.1736	0.0868	0.0868
9	1	2	19.5	0.1026	0.0513	0.1486	0.1000	0.0486
10	17	0	8.5	0.0000	0.0000	0.0000	—	—

习　题

1. 来自 10 份保单的赔付额数据如下：2、3、3、5、5 +、6、7、7 +、9、11 +（+表示损失额超过保单限额，以下同）。使用乘积极限估计，计算保单损失超过 6.5 的概率。

2. 给定含有删失和截断的生存数据如表 8 - 11 所示。

设在时刻 1 后仍未出险的人在时刻 5 或 5 之前出险的概率是${}_3q_1$，用 Greenwood 近似公式估计${}_3\hat{q}_1$的方差。

3. 给定含有删失和截断的生存数据如表 8 - 12 所示。

表 8 - 11

时间（t）	t 时刻的风险数 r_i	t 时刻出险数 s_i
1	30	5
2	27	9
3	32	6
4	25	10
5	27	7

表 8 - 12

时间（t）	t 时刻的风险数 r_i	t 时刻出险数 s_i
1	30	5
2	27	9
3	32	6
4	25	10
5	27	7

使用 Nelson-Åalen 估计来计算 $H(3)$ 的 90% 置信水平的对数变换置信区间。

4. 在两个国家的保险产品的死亡率研究中，给定数据如表 8 - 13 所示。

表 8 - 13

	A　国		B　国	
t_j	S_j	r_j	S_j	r_j
1	20	200	15	100
2	54	180	20	85
3	14	126	20	65
4	22	112	10	45
5	15	110	10	40

其中 r_i 是（t_{i-1}，t_i）期间的风险数，S_i 是（t_{i-1}，t_i）期间的死亡数，并假定全部在 t_i 时刻发生。令 $S^T(t)$ 表示基于两国汇总数据的 $S(t)$ 的乘积极限估计，$S^B(t)$ 是仅基于 B 国数据的 $S(t)$ 的乘积极限估计。求 $|S^T(5) - S^B(5)|$。

5. 累积危险率 $H(t_0)$ 的 95% 置信水平的线性置信区间是（1.63，

1.99)，求其90%置信水平的对数变换置信区间。

6. 一份保单组合产生了如下赔付：100、150、196、250、300、300、400、450、590、770。求累积危险率 $H(300)$ 的经验估计。

7. 对一份保单组合有如下信息，求 $\hat{S}_1(1\ 250)$ 与 $\hat{S}_2(1250)$ 之间差的绝对值：

(1) 各保单都没有免赔额，且保单限额各不相同；

(2) 一个有10个赔付额的样本如下：350、350、500、500、500+、1000、1000+、1000+、1200、1500；

(3) $\hat{S}_1(1\ 250)$ 是 $\hat{S}(1\ 250)$ 的乘积极限估计；

(4) 假设损失额服从指数分布，$\hat{S}_2(1\ 250)$ 是对 $S(1\ 250)$ 的最大似然估计。

8. 来自10份保单的赔付额数据如下：4、4、5+、6+、7+、8、10+、10+、13、15。用Greenwood近似公式估计乘积极限估计 $\hat{S}(12)$ 的方差。

9. 对生存研究中的第 i 个观测，记 d_i 是左截断点，x_i 是没有右删失时的观测值，u_i 是右删失时的观测值。给定表8-14，利用以上数据求 $S_{10}(1.6)$ 的乘积极限估计。

表8-14

观测 (i)	d_i	x_i	u_i	观测 (i)	d_i	x_i	u_i
1	0	0.9	—	6	0	1.7	—
2	0	—	1.2	7	0	—	1.7
3	0	1.5	—	8	1.3	2.1	—
4	0	—	1.5	9	1.5	2.1	—
5	0	—	1.6	10	1.6	—	2.3

10. 某个死亡率研究中包含 n 个人。假设没有删失数据，死亡不会同时发生。已知累计危险率 $\hat{H}(t_2)$ 的Nelson-Åalen估计是59/870。令 t_k 是第 k 次死亡发生的时间，用乘积极限估计求 t_9 时的生存函数值。

11. 若用 $S(t_0)$ 的乘积极限估计 $\hat{S}(t_0)$ 来求其置信区间，已知 $S(t_0)$ 的97.5%置信水平的对数变换置信区间是（0.695，0.777），求 $\hat{S}(t_0)$。

12. 在研究保单退保的研究中，给定下面100份保单的信息，求 n：

(1) 每当有一份保单退保，就会有一份新保单加进来，即风险集 r_j 恒等于100；

(2) 退保只发生在保单年末，其中每个保单年末的退保数如下：第1

年末1份，第2年末2份，第3年末3份…第n年末n份；

（3）时刻n时的累积分布函数的Nelson-Åalen估计为$\hat{F}(n)=0.5975$。

13. 在索赔赔付次数的研究中，假定数据没有删失或截断，一次索赔至多支付一次。已知第二次赔付后的瞬间，累积危险率的Nelson-Åalen估计是17/72。求第四次赔付后的瞬间，累积危险率的Nelson-Åalen估计。

14. 已知（0.357，0.521）是t时刻Nelson-Åalen估计的累积危险率的90%置信水平的对数变换置信区间，求$S(t)$的Nelson-Åalen估计。

15. 一个损失样本中包含以下15个损失数据：11、22、22、22、36、51、69、69、69、92、92、120、161、161、230。设$\hat{H}_1(x)$是累积危险率的Nelson-Åalen经验估计，$\hat{H}_2(x)$是假设样本来自损失服从指数分布的总体时，累积危险率的最大似然估计。求$|\hat{H}_2(75)-\hat{H}_1(75)|$。

16. 12位投保人自保单生效伊始就开始接受观察，直到发生第一次索赔，如表8－15所示。使用Nelson-Åalen估计计算累积危险率$H(4.7)$的90%置信水平的线性置信区间。

表8－15

首次索赔发生时间	1	2	3	4	5	6	7
索赔数	1	3	1	2	1	2	2

17. 在一项生存研究中，死亡发生时间依次为$y_1<y_2<\cdots<y_9$。已知y_6和y_7时刻的累计危险率的Nelson-Åalen估计分别为$\hat{H}(y_6)=0.4128$和$\hat{H}(y_7)=0.5691$，其估计量方差分别$\widehat{Var}(\hat{H}(y_6))=0.009565$和$\widehat{Var}(\hat{H}(y_7))=0.014448$，求$y_7$时的死亡数。

18. 已知数据集$\{y_i\}$为：200、300、100、400、X。给定（1）$k=4$；（2）$s_2=1$；（3）$r_4=1$；（4）Nelson-Åalen估计$\hat{H}(410)>2.15$。求X的值。

19. 已知数据集$\{y_i\}$为：2500、2500、2500、3617、3662、4517、5000、5000、6010、6932、7500、7500。设$\hat{H}_1(7000)$是数据集$\{y_i\}$中没有删失数据时的Nelson-Åalen估计。$\hat{H}_2(7000)$是数据集$\{y_i\}$中仅2500，5000，7500这7个数是右删失时的Nelson-Åalen估计。计算$|\hat{H}_1(7000)-\hat{H}_2(7000)|$。

20. 一个研究右截断数据的死亡率研究中，给定以下数据（见表8－16）。已知时刻为10时生存函数的Nelson-Åalen估计是0.6133，求k的值。

21. 有5位患者从发病到死亡的时间的数据如下：2、3、3、7、8，用带宽为1的三角核函数估计时间为2.5时的密度函数。

22. 设某总体的分布函数是 $F(x)$，给定下列样本数据：2.0、3.3、3.3、4.0、4.0、4.7、4.7、4.7，使用带宽为 1.4 的均匀核函数计算 $F(4)$ 的核密度估计。

23. 来自某一总体的 10 个样本数据为：1、2、3、3、3、3、3、3、3、3。分别记 $\hat{F}_1(x)$ 是使用带宽为 1 的均匀核函数的核密度估计，$\hat{F}_2(x)$ 是使用带宽为 1 的三角核函数的核密度估计。求在 [0, 4] 内满足 $\hat{F}_1(x)=\hat{F}_2(x)$ 的区间。

24. 下面是 10 个观察者的死亡年龄：38、40、46、46、48、50、56、58、60、62，使用带宽为 10 的均匀核函数，求活过 51 岁的概率的核密度估计。

25. 在双重减因模型的研究中，假定个人生存数据受两种减因影响（见表 8－17）。

表 8－16

时刻	死亡数	风险数
3	1	55
5	2	47
6	5	k
10	6	20

表 8－17

序号	年龄	$q_j^{(T)}$
0	0	0.150
1	30	0.196
2	60	0.727
3	90	1.000

已知：（1）$q_j^{'(2)}=0.06$ 对所有 j 成立；（2）A 组包含 1 000 组数据，观测期从 0 开始；（3）A 组仅被第一种减因影响。求 A 组至少能活到 60 岁的人数的 Kaplan－Meier 多重损因估计。

第九章　参数模型的估计

学习目标

- □ 掌握完整样本数据下个体数据和分组数据的矩估计、分位数估计和极大似然估计方法
- □ 掌握非完整样本数据（存在删失和截断的数据）的矩估计和极大似然估计方法
- □ 熟悉极大似然估计量方差的计算，了解 Fisher 信息量的含义
- □ 了解多变量参数模型的特点，熟悉二元变量模型、和模型、Cox 模型、广义线性模型等多变量参数模型的参数估计

§ 9.1　完整数据情况下参数的点估计

9.1.1　矩估计

假设 n 个样本观测值（x_1，…，x_n）来自于同一个总体。总体分布函数为：

$$F(x,\theta),\quad \theta^T=(\theta_1,\theta_2,\cdots,\theta_p)$$

其中 θ^T 表示 θ 的转置，θ 为列向量，含有 p 个待估参数。进一步，令 $\mu'_k(\theta)=E(X^k\mid\theta)$ 表示 k 阶原点距，$\mu_k(\theta)=E((X-E(X))^k\mid\theta)$ 表示 k 阶中心距。矩估计方法的基本思想是求解参数 θ 使得样本分布的前 p 阶原点矩等于经验估计的前 p 阶原点矩，即

$$\mu'_k(\theta)=\mu'_k,\quad k=1,\cdots,p \tag{9.1.1}$$

有时也可以用中心矩来求参数 θ，即

$$\mu_k(\theta)=\mu_k,\quad k=1,\cdots,p \tag{9.1.2}$$

对于个体完整数据，样本矩的计算和矩估计的原理在数理统计教材中都有详细论述，本章将不再进行叙述。对于分组数据，由它的光滑经验分布曲线可计算 k 阶样本原点矩为：

$$\hat{\mu}'_k=\sum_{j=1}^{r}\int_{c_{j-1}}^{c_j}x^k\frac{n_j}{n(c_j-c_{j-1})}dx=\sum_{j=1}^{r}\frac{n_j(c_j^{k+1}-c_{j-1}^{k+1})}{n(c_j-c_{j-1})(k+1)} \tag{9.1.3}$$

【例 9－1】 假设用伽玛分布来拟合例 8－1 的数据，试用矩估计方法计算相应的分布参数。

解：对于伽玛分布，$f_X(x)=\dfrac{1}{\theta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\theta}$，矩估计方程为：

$$E(X)=\alpha\theta=22\ 626.5$$

$$E(X^2)=\alpha(\alpha+1)\theta^2=2\ 719\ 971\ 502$$

解得：$\hat{\alpha}=0.232$，$\hat{\theta}=97\ 585.27$。■

【例9-2】 假设某险种的随机理赔数据分布如表9-1所示，假设理赔额服从帕累托分布，$f(x)=\dfrac{\alpha\theta^{\alpha}}{(\theta+x)^{\alpha+1}}$，试用矩估计法估计两个参数 α 和 θ。

表9-1　某险种随机理赔数据

理赔额范围	理赔数
0～50	4
50～100	7
100～200	9
200～400	16
400～1 000	29
1 000～2 000	12
2 000～5 000	3
合计	80

解： 由光滑经验分布曲线计算样本原点矩为：

$$\begin{aligned}\hat{\mu}' &= \sum_{j=1}^{7}\frac{n_j(c_j^2-c_{j-1}^2)}{(80)(2)(c_j-c_{j-1})}\\ &= \sum_{j=1}^{7}\frac{n_j(c_j+c_{j-1})}{(80)(2)}\\ &= \left(\frac{4}{80}\right)(25)+\left(\frac{7}{80}\right)(75)+\left(\frac{9}{80}\right)(150)+\left(\frac{16}{80}\right)(300)\\ &\quad+\left(\frac{29}{80}\right)(700)+\left(\frac{12}{80}\right)(1\ 500)+\left(\frac{3}{80}\right)(3\ 500)=694.69\end{aligned}$$

$$\begin{aligned}\hat{\mu}_2' &= \sum_{j=1}^{7}\frac{n_j(c_j^3-c_{j-1}^3)}{n(3)(c_j-c_{j-1})}=\frac{1}{240}\left[\frac{4(50^3-0^3)}{50-0}+\cdots+\frac{3(5\ 000^3-2\ 000^3)}{5\ 000-2\ 000}\right]\\ &=1\ 047\ 843.75\end{aligned}$$

对于帕累托分布，两个矩估计方程为：

$$E(X)=\frac{\theta}{\alpha-1}=694.69$$

$$E(X^2)=\frac{2\theta^2}{(\alpha-1)(\alpha-2)}=1\ 047\ 843.75$$

解得：$\hat{\alpha}=13.7$，$\hat{\theta}=8\ 806.8$。■

9.1.2　分位数估计

矩估计的基本思想是求取参数 θ 使分布的矩与样本的矩相匹配。分位数估计法则是另一种匹配法，其基本思想是构造一个模型使其 p 个分布分位点与实际样本的 p 个分位点相匹配。

定义9-1　假定 X 服从分布函数为 $F(x)$ 的分布，令 $0<p<1$，满足等式 $F(\pi_p)=P(X\leqslant\pi_p)=p$ 的 π_p 称为分布 $F(x)$ 的 p 分位数。当 $p=50\%$ 时，$\pi_{0.5}$ 称为**中位数**。

当 X 的分布是连续分布时，p 分位数是唯一确定的。但是当 X 的分布是离散分布时，p 分位点并不总是有明确的定义，不能唯一确定。例如，X

取值于 0 和 1 的概率均为 1/2，对于 $p=0.6$ 无法确定分位数 $\pi_{0.6}$。因此分位数估计只能适用于连续分布的参数估计。

当样本数据是个体完整数据时，其经验分布函数是离散的分布，使用分位数的一个问题在于样本 p 分位点的确定。比如例 9－1 的数据中，在7 947～8 391 之间的任何一个数，都满足有 10 个观测值大于它、同时又有 10 个观测值小于它，因此都可以作为中位数，但习惯上我们使用这个区间的中点。对于其他 p 分位点，样本的经验分位点通常采用插值的方法获得。如果随机变量 X 有 n 个观测值 $x_1, x_2, \cdots, x_n$，将这些观测值从小到大排列为 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$，对于 $0<p<1$，$g=[(n+1)p]$表示不超过$(n+1)p$ 的最大整数，此时认为分位数应该在 $x_{(g)}$ 和 $x_{(g+1)}$之间。记 $h=(n+1)p-g$ 表示$(n+1)p$ 的小数部分，则定义样本的 $100p\%$ 的分位数如下：

定义 9－2 个体数据的 $100p\%$ 分位点的光滑经验估计为：

$$\hat{\pi}_p = (1-h)x_{(g)} + hx_{(g+1)} \tag{9.1.4}$$

当 $p=50\%$ 时，$\hat{\pi}_{0.5}$称为**样本中位数**。当 n 为奇数时，记 $k=(n+1)/2$，样本中位数为 $x_{(k)}$，当 n 为偶数时，记 $k=n/2$，则样本中位数为$\frac{1}{2}x_{(k)}+\frac{1}{2}x_{(k+1)}$。样本中位数是位置平均，不受极大值和极小值的影响，主要用于测度顺序数据的集中趋势。

注意，只有当样本数据中没有重复的观测值时，定义 9－2 中任何两个分位点的值都不相同。另外，我们无法获得 $p<1/(n+1)$ 和 $p>n/(n+1)$ 情况下的值。这个结论也是合理，因为对于小样本我们不应期望能够得到较大或者较小的分位点。

对于分组数据，根据其经验分布公式

$$F_n(x) = \begin{cases} 0, & x \leqslant c_0 \\ \dfrac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \dfrac{(x - c_{j-1})}{c_j - c_{j-1}} F_n(c_j), & c_{j-1} \leqslant x < c_j \\ 1, & x > c_r \end{cases}$$

其中 $F_n(c_j) = \frac{1}{n}\sum_{i=1}^{j} n_i$。对于 $0<p<1$，分组数据的样本 $100p\%$ 的分位点 $\hat{\pi}_p$ 定义为：

$$p = F_n(\hat{\pi}_p) \tag{9.1.5}$$

即
$$\hat{\pi}_p = \frac{p - F_n(c_{j-1})}{F_n(c_j) - F_n(c_{j-1})} \times c_j + \frac{F_n(c_j) - p}{F_n(c_j) - F_n(c_{j-1})} \times c_{j-1} \tag{9.1.6}$$

或者
$$\hat{\pi}_p = c_{j-1} + \left(np - \sum_{i=1}^{j-1} n_i\right) \frac{c_j - c_{j-1}}{n_j} \tag{9.1.7}$$

其中，c_{j-1}和 c_j 满足 $F_n(c_{j-1}) \leqslant p < F_n(c_j)$。

定义 9－3　分位数估计是指满足下列 p 个方程的任意解：

$$\pi_{g_k}(\theta)=\hat{\pi}_{g_k},\ k=1,\ 2,\ \cdots,\ p \tag{9.1.8}$$

其中，g_1，g_2，…，g_p 是 p 个的百分数。

当只有一个待估参数时，通常选取 50% 的分位数来估计参数；当有两个待估参数时，通常选取 25% 和 75% 的分位数来估计。由分位点的定义可知，方程（9.1.8）也可以写成

$$F(\hat{\pi}_{g_k}\mid\theta)=g_k,\qquad k=1,\ 2,\ \cdots,\ p$$

【例 9－3】　设表 9－2 中的火灾损失额服从韦伯分布，用 25% 和 75% 分位数来估计参数的值。

表 9－2　　火灾损失数据

0.1	0.5	2.2	4.1	28.1
0.2	0.7	2.6	5.9	30.0
0.2	0.9	2.9	6.2	49.2
0.3	1.3	3.2	12.1	63.8
0.4	1.8	3.3	13.65	118.0

解：韦伯的分布函数为：

$$F_X(x)=1-e^{-x^\gamma/\theta}$$

分别令 $0.75=1-e^{-x^\gamma/\theta}$，$0.25=1-e^{-x^\gamma/\theta}$，解得：

$$x_{0.75}=(\theta\ \ln 4)^{\frac{1}{\gamma}},$$

$$x_{0.25}=(\theta\ \ln\ (4/3))^{\frac{1}{\gamma}}$$

由于 $0.25\times26=6.5$，因此，

0.25 的分位点为：$0.5\times0.5+0.5\times0.7=0.6$

0.75 的分位点为：$0.5\times12.1+0.5\times13.65=12.875$

类似地，由分位数估计法 $12.875=(\theta\ \ln 4)^{\frac{1}{\gamma}}$，$0.65=(\theta\ \ln\ (4/3))^{\frac{1}{\gamma}}$ 解得：$\gamma=0.5129$，$\theta=2.675$。从而确定了损失分布。■

【例 9－4】　由 10 只实验老鼠组成的样本，其死亡时间（以天为单位）为 3、4、5、7、7、8、10、10、10、12。假设适合的生存模型服从 Gompertz 分布，利用 25% 和 65% 分位点，估计参数 B 和 c。

解：由 $n=10$，$0.25\times11=2.75$，$0.65\times11=7.15$，有

$$\pi_{0.25}=0.25\times X_{(2)}+0.75\times X_{(3)}=4.75$$

$$\pi_{0.65}=0.85\times X_{(7)}+0.15\times X_{(8)}=10$$

Gompertz 分布的分布函数为：

$$F(x)=e^{-\int_0^x h(y)dy}=1-e^{\frac{B}{\ln c}(1-c^x)},x>0,B>0,c\geqslant1$$

因此，$\pi_{0.25}=\dfrac{\ln\left[1-\dfrac{\ln(0.75)\ln c}{B}\right]}{\ln c}$，　$\pi_{0.65}=\dfrac{\ln\left[1-\dfrac{\ln(0.35)\ln c}{B}\right]}{\ln c}$

令 $\pi_{0.25}=4.75$，$\pi_{0.65}=10$，则由所给等式解得 $\hat{B}=0.039$，$\hat{c}=1.1896$。■

【例 9－5】　某责任险保单规定了保单限额为 30 万元，表 9－3 中的第 1 至 3 列给出了该险种 217 份保单的理赔额情况。假设理赔额服从对数正态分布，请用 30% 和 70% 分位数来估计参数。

表 9－3　　　　某责任险理赔情况

理赔额	保单数	平均理赔额
0～2 500	41	1 389
2 500～7 500	48	4 661
7 500～12 500	24	9 991
12 500～17 500	18	15 482
17 500～22 500	15	20 232
22 500～32 500	14	26 616
32 500～47 500	16	40 278
47 500～67 500	12	56 414
67 500～87 500	6	74 985
87 500～125 000	11	106 851
125 000～225 000	5	184 735
225 000～300 000	4	264 025
300 000	3	300 000

解： 经验分布 30% 和 70% 的分位点分别为：

$$2\ 500 + (65.1 - 41) \times 5\ 000/48 = 5\ 010$$

$$22\ 500 + (151.9 - 146) \times 10\ 000/14 = 26\ 714$$

由于理赔额服从对数正态分布，因此 $\ln X \sim N(\mu, \theta)$，令 5 010 和 26 714为对数正态分布的分位点，即

$$0.3 = \Phi[(\ln 5\ 010 - \mu)/\sigma]$$

$$0.7 = \Phi[(\ln 26\ 714 - \mu)/\sigma]$$

解方程组得：$\hat{\sigma} = 1.595871$，$\hat{\mu} = 9.356065$。 ■

9.1.3　极大似然估计

设数据集由 n 个事件 A_1，…，A_n 组成。如果是个体数据，则 A_j 为第 j 个观测，它是通过观测随机变量 X_j 得到的。如果是分组数据或截断数据则 A_j 可能是一个区间，例如在 u 处删失的观测可能代表发生在 u 到正无穷区间的事件。在这里无需要求 X_1，…，X_n 具有相同的分布，但是要求其分布依赖同一个参数向量 θ，而且这些随机变量要相互独立。

定义 9－4　似然函数的定义为：

$$L(\theta) = \prod_{j=1}^{n} \mathrm{P}(X_j \in A_j \mid \theta) \tag{9.1.9}$$

并且，θ 的极大似然估计是使似然函数取最大值的向量。

当数据没有截断也没有删失，并且每一个点都被记录的时候，很容易写出似然函数和对数似然函数：

$$L(\theta) = \prod_{j=1}^{n} f_{X_j}(x_j \mid \theta), \quad l(\theta) = \sum_{j=1}^{n} \ln f_{X_j}(x_j \mid \theta) \tag{9.1.10}$$

注意这个表达式并不要求观测都来自同一个分布。对于个体完整数据的极大似然估计，在一般的统计教材中都有详细论述，我们在此不再叙述。下面的例子是极大似然估计在生存模型中的应用。

【例 9－6】 对于由 5 个从出生起便受到伤害的个体组成的样本，假定死亡力为年龄的线性函数，即 $\mu_x = bx$，如果死亡发生时间为 1，2，3，4，5，求 b 的极大似然估计量。

解： 根据死亡力函数可以推出生存函数：

$$S_X(x) = \exp\left[-\int_0^x h(s)ds \right] = e^{-\int_0^x bsds} = e^{-\frac{b}{2}x^2}$$

进而推出分布函数和密度函数：

$$F_X(x) = 1 - S_x(x) = 1 - e^{-\frac{b}{2}x^2}, f_X(x) = bxe^{-\frac{b}{2}x^2}$$

根据死亡发生时间写出似然函数：

$$L(b) = \prod_{i=1}^{5} f_X(x_i) = b^5 \prod_{i=1}^{5} x_i e^{-\frac{b}{2}\sum_{i=1}^{5} x_i^2} = 120b^5 e^{-27.5b}$$

对数似然函数为：

$$l(b) = 5\ln b - 27.5b + \ln 120$$

令 $l'(b) = 0$，解得：$b = 2/11$。 ■

当数据是完全的并且是分组形式时，可以对观测值进行如下归纳：首先选取一组数 $c_0 < c_1 < \cdots < c_k$，这里 c_0 是最小的可能观测值（通常为 0），c_k 是最大的可能观测值（通常为正无穷）。设 n_j 为落入区间 $(c_{j-1}, c_j]$ 的观测数目，这种数据的似然函数为：

$$L(\theta) = \prod_{j=1}^{k} [F(c_j \mid \theta) - F(c_{j-1} \mid \theta)]^{n_j} \tag{9.1.11}$$

其对数为：

$$l(\theta) = \sum_{j=1}^{k} n_j \ln[F(c_j \mid \theta) - F(c_{j-1} \mid \theta)] \tag{9.1.12}$$

【例 9－7】 考察一个在 $t = 0$ 处有 20 个个体的样本，所有的个体均在 5 周内死亡，并只记录每周的死亡人数，所观察的结果如下：2 人在第 1 周死亡；3 人在第 2 周死亡；8 人在第 3 周死亡，6 人在第 4 周死亡，1 人在第 5 周死亡。假设适合模型为指数模型，求参数 θ 的极大似然估计。

解： 画图来帮助理解，见图 9－1。

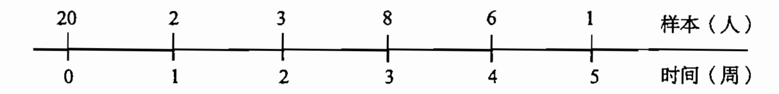


图 9－1

对于参数为 θ 的指数模型，$S(x)=e^{-x/\theta}$，那么在第 i 周死亡的概率为：

$$S(i-1)-S(i)=e^{-(i-1)/\theta}-e^{-i/\theta}=e^{-i/\theta}(e^{1/\theta}-1)$$

所以似然函数

$$L(\theta)=\prod_{i=1}^{5}[e^{-i/\theta}(e^{1/\theta}-1)]^{n_i}$$

$$l(\theta)=-\frac{1}{\theta}\sum_{i=1}^{5}i\cdot n_i+\ln(e^{1/\theta}-1)\cdot\sum_{i=1}^{5}n_i$$

令 $\dfrac{d\ln L}{d\theta}=\dfrac{1}{\theta^2}\displaystyle\sum_{i=1}^{5}i\cdot n_i-\dfrac{1}{\theta^2}\dfrac{e^{1/\theta}\cdot\sum_{i=1}^{5}n_i}{e^{1/\theta}-1}=0$，由此得：

$$e^{1/\theta}=\frac{\sum_{i=1}^{5}i\cdot n_i}{\sum_{i=1}^{5}i\cdot n_i-\sum_{i=1}^{5}n_i}$$

由 $n_1=2$，$n_2=3$，$n_3=8$，$n_4=6$，$n_5=1$，于是有 $\sum_{i=1}^{5}n_i=20$，$\sum_{i=1}^{5}i\cdot n_i=61$，因此，$\hat{\theta}=2.5170$。■

【例9-8】 假设某险种的理赔数据如表9-4，若假设理赔额服从指数分布，试用极大似然估计求指数分布的参数 θ。

解：若假设理赔额服从指数分布，根据式（9.1.12），其对数似然函数为：

$$\begin{aligned}l(\theta)&=99\ln[F(7\,500)-F(0)]+42\ln[F(17\,500)-F(7\,500)]+\cdots\\&\quad+3\ln[1-F(300\,000)]=99\ln[1-e^{-7\,500/\theta}]\\&\quad+42\ln[e^{-7\,500/\theta}-e^{-17\,500/\theta}]+\cdots+3\ln e^{-300\,000/\theta}\end{aligned}$$

使用数值算法得到极大似然估计为 $\hat{\theta}=29\,721$，似然函数值为 -406.03。■

表9-4　某险种理赔数据

理赔额范围	0 ~ 7 500	7 500 ~ 17 500	17 500 ~ 32 500	32 500 ~ 67 500	67 500 ~ 125 000	125 000 ~ 300 000	300 000 +
理赔数	99	42	29	28	17	9	3

【例9-9】 表9-5给出了某医疗责任保单组在10年内的发生的年索赔数。假设泊松分布和负二项分布都适合描述这张保单组的年索赔数分布，使用极大似然法估计泊松参数和负二项分布参数。

表9-5　某医疗责任险年索赔数

年　份	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
索赔数	6	2	3	0	2	1	2	5	1	3

解： 可以按照不同的方法对这些数据进行综合。例如只有 1988 年这一年没有发生索赔，1990 年、1993 年这两年都只发生了一次索赔。统计结果见表 9－6。

表 9－6　　医疗责任险年索赔频数

频数（k）	0	1	2	3	4	5	6	7＋
观测数（n_k）	1	2	3	2	0	1	1	0

在 1985—1994 间的索赔总数为 25。因此，平均每年的索赔次数为 2.5，这个平均值也可由表 9－6 算出。令 n_k 表示恰好出现 k 次索赔的年数，发生 k 次索赔的概率为 p_k，则整个观测集的似然函数和对数似然函数分别为：

$$L = \prod_{k=0}^{\infty} p_k^{n_k}, l = \sum_{k=0}^{\infty} n_k \ln p_k$$

对于泊松分布，有

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}, \ln p_k = -\lambda + k \ln \lambda - \ln k!$$

对数似然函数为：

$$l = \sum_{k=0}^{\infty} n_k(-\lambda + k \ln \lambda - \ln(k!)) = -\lambda n + \sum_{k=0}^{\infty} kn_k \ln \lambda - \sum_{k=0}^{\infty} n_k \ln(k!)$$

其中 $n = \sum_{k=0}^{\infty} n_k$ 为样本容量。对数似然函数关于 λ 求微分，得到：

$$\frac{dl}{d\lambda} = -n + \sum_{k=0}^{\infty} kn_k \frac{1}{\lambda}$$

令对数似然函数的导数值为零，求解后可以得到极大似然估计量。估计量为：

$$\hat{\lambda} = \frac{\sum_{k=0}^{\infty} kn_k}{n} = \bar{x}$$

对于负二项分布，有

$$p_k = \binom{r+k-1}{k} p^r q^k, \quad 0 < p < 1, p + q = 1, k = 0, 1, 2, \cdots$$

其对数似然函数为两个参数 p 和 r 的函数：

$$l = \sum_{k=0}^{\infty} n_k\left[\ln \frac{(r+k-1)!}{k!(r-1)!} + k \ln(1-p) + r \ln p\right]$$

为了求对数似然函数的最大值，分别对每个参数求导并令导数为零，然后解参数。对数似然函数的导数为：

$$\frac{\partial l}{\partial p} = \sum_{k=0}^{\infty} n_k\left(\frac{r}{p} - \frac{k}{1-p}\right)$$

$$\frac{\partial l}{\partial r} = \sum_{k=0}^{\infty} n_k \ln p + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \ln \frac{(r+k-1)\cdots r}{k!}$$

$$= n \ln p + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \sum_{m=0}^{k-1} \ln(r+m) = n \ln p + \sum_{k=1}^{\infty} n_k \sum_{m=0}^{k-1} \frac{1}{r+m}$$

令导数为零，解得：

$$\hat{r}\frac{1-\hat{p}}{\hat{p}} = \bar{x} = \frac{1}{n}\sum_{k=0}^{j} kn_k \tag{9.1.13}$$

$$\ln\left(\frac{1}{p}\right) = \frac{1}{n}\sum_{k=0}^{j} n_k\left(\sum_{m=0}^{k-1}\frac{1}{\hat{r}+m}\right) \tag{9.1.14}$$

式(9.1.13)和(9.1.14)可以用数值方法求解。将 $\hat{p}$ 用 $\hat{r}/(\bar{x}+\hat{r})$ 替换得到：

$$\hat{H}(r) = n \ln\left(1+\frac{\bar{x}}{\hat{r}}\right) - \sum_{k=1}^{\infty}\left(\sum_{m=0}^{k-1}\frac{1}{\hat{r}+m}\right) = 0 \tag{9.1.15}$$

如果样本方差大于样本均值，则可以证明(9.1.15)存在唯一解。

用 Newton-Raphson 方法对方程(9.1.15)求数值解，解得 $\hat{r}=10.965$ 和 $\hat{p}=0.186$。 ■

§ 9.2 非完整数据情况下参数的点估计

对于非完整数据，本节将运用极大似然估计法讨论参数模型的估计。虽然矩估计法和分位点匹配法也可以用来估计参数，而且通常来说比较容易操作，但是这些方法还是有一定缺陷的。首先这些方法只用到了数据的部分性质，没有能够充分地利用全部数据。特别是当总体分布的右尾很厚时，是否能够尽可能充分利用数据的信息显得尤为重要。例如，对于正态分布的参数估计，样本均值和方差提供的信息将能够完全代替原始数据。但是，在估计帕累托分布的参数时，为了正确估计参数 α，需要掌握所有的极值观测；而当数据存在删失情况时，这些数据的极值观测是无法获得的。其次，这些方法要求所有的观测来自同一个随机变量，否则很难确定如何用数据来描述总体的矩和分位点。例如，当一半的观测来自免赔额为 100 的业务，而另一半观测来自免赔额为 200 的业务时，显然全部样本的样本均值是没有确切含义的。最后，这些方法允许人们任意选取矩的阶数和分位数。

根据非完整数据存在右删失和左截断的情况，我们分别讨论以下几种情况的极大似然估计。

9.2.1 右删失数据的极大似然估计

设 $x_1, x_2, \cdots, x_n$ 是随机变量 X 的 n 个观测，将它们从小到大排列 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$。假设样本数据在 $x=u$ 处存在右删失，且存在 $1 \leqslant k < n$ 使得 $x_{(k)} < u \leqslant x_{(k+1)}$，则样本数据将变为 $x_{(1)}, x_{(2)}, \cdots, x_{(k)}, u, \cdots u$。似然函数可以写为：

$$L(\theta) = \left[\prod_{i=1}^{k} f(x_{(i)};\theta)\right] S(u;\theta)^{n-k} \tag{9.2.1}$$

【例 9-10】 承例 9-4，由 10 只实验老鼠组成的样本，假设生存模型服从指数分布，试验样本的死亡时间（以天为单位）为 3、4、5、7、7、8、10、10、10、12。假设研究在 $t=9$ 处截尾，在这种情况下估计指数参数 θ。

解：因为研究在 $t=9$ 处截尾，所以数据样本是右删失的；在 $t=9$ 处，已经有 6 只死亡，有 4 只仍然生存，于是根据式(9.2.1)有

$$L = \prod_{i=1}^{6} f(t_i)[S(9)]^4$$

根据指数分布的性质有：

$$S(9) = e^{-9/\theta}, f(t_i) = \frac{1}{\theta}e^{-t_i/\theta}, \quad t_i = 3,4,5,7,7,8$$

由此可以得到似然函数：$L = \theta^{-6}e^{-70/\theta}$，则 $\ln L = -6\ln\theta - 70/\theta$。对数似然函数求导得 0 解得 $\hat{\theta} = 70/6 = 11.67$。 ■

【例 9-11】 已知某险种的赔偿限额为 50，随机抽取的 12 个理赔额分别为：3，4，8，10，12，18，22，35，50，50，50，50。假设实际损失额 X 服从在 $[0, \theta]$ 上的均匀分布，利用极大似然方法估计出参数 θ。

解：因为赔偿限额为 50，所以理赔数据在 50 处右删失；根据式(9.2.1)，在赔偿限额为 50 的条件下，似然函数为：

$$L(\theta) = \prod_{i=1}^{8} f(x_i;\theta)(S(50))^4$$

在均匀分布的假设下，

$$S(50) = P[X > 50] = \frac{\theta - 50}{\theta}$$

于是似然函数为：

$$L(\theta) = \prod_{i=1}^{8} f(x_i;\theta)(P[X > 50])^4 = \left(\frac{1}{\theta}\right)^8\left(\frac{\theta - 50}{\theta}\right)^4 = \frac{(\theta - 50)^4}{\theta^{12}}$$

令 $\dfrac{dl(\theta)}{d\theta} = \dfrac{4}{\theta - 50} - \dfrac{12}{\theta} = 0$，解得 $\hat{\theta} = 75$。 ■

9.2.2 左截断数据的极大似然估计

当数据存在左截断时，截断点 d 下方数据将不会被观测。例如假设保单具有免赔额 d，那么意味着小于 d 的实际损失额将不会被报告。截断数据有两种记录方式。如果数据只被截断，则截断后的观测值为：

$$Y = X \mid X > d \tag{9.2.2}$$

称 Y 为在 d 点截断数据。若 X 表示实际损失额（死亡时间），则 Y 即为高于 d 的损失（死亡时间）。

另一种记录方式是记录超过截断点 d 的部分值，这时截断后的观测值为：

$$W = (X - d) \mid X > d \tag{9.2.3}$$

称 W 为截断且被平移的数据。若 d 表示免赔额，X 表示损失事件的实际损失额，则 W 就是每次理赔事件的理赔额。当实际损失小于免赔额 d 时，被保险人没有获得理赔，理赔额不存在，因而没有定义。

设总体分布为 $F(x, \theta)$，x_1，…，x_n 为没有被截断的观测值。假设样本数据在 $x = d$ 处存在左截断，且存在 $1 \leqslant k < n$ 使得 $x_{(k-1)} \leqslant d < x_{(k)}$。如果只记录 d 点截断数据，则样本数据将变为 $x_{(k)}$，$x_{(k+1)}$，…，$x_{(n)}$。如果记录的是截断且被平移的数据，则样本数据为 $x_{(k)} - d$，$x_{(k+1)} - d$，…，$x_{(n)} - d$，用 w_1，…，w_{n-k}来表示。

对于这些被截断的数据，一般有两种处理办法：一种是对数据平移，即将每个观测点减去截断点；另一种方法是不进行平移，认为数据下方的数据对模型的拟合没有带来任何信息。对截断点上方的数据设置条件概率，似然函数中的概率为条件概率。具体来说：

1. 平移法（*The shifted model*）。对于被截断且被平移的观测值，由于这些观测值都是从 O 开始被记录，从数据本身来看，就像没有被截断的数据。假设被截断且被平移的观测值(w_1，w_2，…，w_k)具有与没有被截断的样本相同的分布类型，只是参数不同，那么就可以不考虑截断点对数据的影响直接使用分布进行估计。比如，若 d 表示免赔额，假设实际损失 X 的分布为 $F(x, \theta)$，若每次理赔事件的理赔额 W 的分布为 $F(x, \theta')$，则可以直接使用理赔额观测值对参数 θ'进行估计。

对于只被截断的观测值，可以先对数据平移，即将每个观测点减去截断点，然后使用平移后分布进行估计。这时，似然函数为：

$$L(\theta) = \prod_{i=k}^{n} f(x_i - d;\theta) = \prod_{i=1}^{n-k} f(w_i;\theta) \tag{9.2.4}$$

【例 9－12】 已知一组的工伤险赔付数据如下：35、72、120、135、165、144、243、302、332、378、465、664、854、858、986、1 185、1 332、1 892、2 567、15 730；假设它在 200 处从下方被截断。使用平移法估计已知 $\theta = 800$ 的帕累托分布的参数 α，并计算当免赔额为 0、200、400 时理赔事件理赔额的期望值。

解：因为数据在 200 处截断，所以被记录下来的样本数据只有 14 个。使用数据平移法，得到截断且被平移的数据为：43、102、132、178、265、464、654、658、786、985、1 132、1 692、2 367、15 530。对于帕累托分布，$f(x) = \dfrac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}}$，因此似然函数为：

$$L(\alpha) = \prod_{j=1}^{14} \frac{\alpha(800^{\alpha})}{(800 + w_j)^{\alpha+1}}$$

对数似然函数：

$$l(\alpha)=\sum_{j=1}^{14}[\ln\alpha+\alpha\ln 800-(\alpha+1)\ln(w_j+800)]$$
$$=14\ln\alpha-10.353\alpha-103.938$$

令 $l'(\alpha)=14/\alpha-10.353=0$，得$\hat{\alpha}=1.352$。

下面计算不同免赔额对应的理赔额期望值：

（1）免赔额等于0。因为数据被移动了，所以无法估计无免赔时的成本。

（2）当免赔额为200时，这些在200处被截断且被平移的数据可看做是每次理赔事件理赔额的样本数据，因此每次理赔事件的理赔额期望值为 $E(W)=\dfrac{\theta}{\alpha-1}$，即 $800/0.352=2\ 272.7$。

（3）当免赔额为400时，等价于在这些截断的数据上再次附加免赔额为200的条款，因此，每次理赔事件的理赔额的期望为：

$$\frac{E(W)-E(W\wedge 200)}{1-F(200)}=\frac{\dfrac{800}{0.352}\left(\dfrac{800}{200+800}\right)^{0.352}}{\left(\dfrac{800}{200+800}\right)^{1.352}}=2\ 840.85 \quad ■$$

2. 非平移法。如果不对数据进行平移，建立似然函数的一个问题是对高于截断点 d 的那些观测值设置条件概率。设 d 为截断点，Y 是被截断后 X 的值，设 X 的分布函数为 $F(x)$，由式（9.2.2）知 Y 的分布为：

$$F_Y(x)=\begin{cases}0, & x\leqslant d\\ P(X\leqslant x\mid X>d), & x>d\end{cases}=\begin{cases}0, & x\leqslant d\\ \dfrac{F(x)-F(d)}{1-F(d)}, & x>d\end{cases} \tag{9.2.5}$$

若 X 是连续随机变量，则 Y 的密度函数为：

$$f_Y(x)=\begin{cases}0, & x\leqslant d\\ \dfrac{f(x)}{1-F(d)}, & x>d\end{cases} \tag{9.2.6}$$

因此被截断后的观测值 y_1，…，y_k 似然函数值为：

$$\prod_{i=1}^{k}\frac{f(y_i;\theta)}{1-F(d;\theta)}$$

设 W 为被截断且被平移后的值，类似地可推出 W 的分布为：

$$F_W(x)=\begin{cases}0, & x\leqslant 0\\ \dfrac{F(x+d)-F(d)}{1-F(d)}, & x>0\end{cases} \tag{9.2.7}$$

$$f_W(x)=\begin{cases}0, & x\leqslant 0\\ \dfrac{f(x+d)}{1-F(d)}, & x>0\end{cases} \tag{9.2.8}$$

因此被截断且被平移后的观测值 w_1，…，w_k 似然函数值为：

$$L(\theta) = \prod_{i=1}^{k} \frac{f(w_i + d;\theta)}{1 - F(d;\theta)}$$

【例 9－13】 承例 9－12，已知一组人造的工伤险赔付数据如下：35、72、120、135、165、144、243、302、332、378、465、664、854、858、986、1 185、1 332、1 892、2 567、15 730；假设它在 200 处从下方截断。不对数据进行平移，使用极大似然估计已知 $\theta=800$ 的帕累托分布的参数 α，并计算当免赔额为 0、200、400 时理赔事件理赔额的期望值。

解： 观察赔付数据可知有 14 个数据大于截断点 200，不对数据进行平移，可以写出似然函数（这里的 x_j 为原始数据）：

$$L(\alpha) = \prod_{j=1}^{14} \frac{f(x_j \mid \alpha)}{1 - F(200 \mid \alpha)} = \prod_{j=1}^{14} \left[\frac{\alpha(800^{\alpha})}{(800 + x_j)^{\alpha+1}} \Big/ \left(\frac{800}{800 + 200}\right)^{\alpha} \right]$$

$$= \prod_{j=1}^{14} \frac{\alpha(1\ 000^{\alpha})}{(800 + x_j)^{\alpha+1}}$$

进而可得到对数似然函数：

$$l(\alpha) = 14 \ln \alpha + 14\alpha \ln 1\ 000 - (\alpha + 1) \sum_{j=1}^{14} \ln(800 + x_j)$$

$$= 14 \ln \alpha + 96.709\alpha - (\alpha + 1)105.76$$

将对数似然函数求导得，$l'(\alpha) = 14\alpha^{-1} - 9.051$，令 $l'(\alpha) = 0$ 解得：$\hat{\alpha} = 1.547$。

当免赔额为 200、400 时，参照例 9－14 得到如下结果：

$$\frac{E(X) - E(X \wedge 200)}{1 - F(200)} = \frac{\dfrac{800}{0.547}\left(\dfrac{800}{200 + 800}\right)^{0.547}}{\left(\dfrac{800}{200 + 800}\right)^{1.547}} = 1\ 828.15$$

$$\frac{E(X) - E(X \wedge 400)}{1 - F(400)} = \frac{\dfrac{800}{0.547}\left(\dfrac{800}{400 + 800}\right)^{0.547}}{\left(\dfrac{800}{400 + 800}\right)^{1.547}} = 2\ 193.78$$

■

【例 9－14】 已知某险种的免赔额为 5，随机抽取 8 个理赔事件的理赔额如下：3、4、8、10、12、18、22、35。假设实际损失额 X 服从指数分布，用极大似然法估计其参数 θ。

解： 因为最终要估计的是实际损失额的参数，所以先将理赔额还原成实际损失额，$x_i = w_i + d$，于是得到的被截断样本数据为 8、9、13、15、17、23、27、40。根据式（9.2.6）写出似然函数：

$$L(\theta) = \prod_{i=1}^{8} \frac{f(x_i;\theta)}{1 - F(d;\theta)} = \frac{\prod_{i=1}^{8}\left(\frac{1}{\theta}e^{-x_i/\theta}\right)}{(e^{-d/\theta})^{8}} = \frac{1}{\theta^{8}} e^{-(\sum x_i - 8d)/\theta}$$

$$= \frac{1}{\theta^{8}} e^{-(152-40)/\theta} = \frac{1}{\theta^{8}} e^{-(112)/\theta}$$

从而可以得到对数似然函数：

$$l(\theta) = -8\ \ln\theta - 112/\theta$$

令 $l'(\theta)=0$，解得：$\hat{\theta}=14$。 ■

【例 9－15】 100 个 x 岁年龄组的人群中，在区间$(x, x+1]$上观察到有 98 人生存到$(x+1)$岁，2 人死亡，死亡分别发生在$(x+0.2)$岁和$(x+0.6)$岁。假设在区间$(x, x+1]$死亡力为常数，求 q_x 的极大似然估计值。

解：第 i 个死亡者的似然函数值由在 x 岁生存条件下、在 x_i 处死亡的密度函数给出：

$$L_i = f(x_i \mid X > x) = \frac{f(x_i)}{S(x)} = \frac{S(x_i)h(x_i)}{S(x)} \tag{9.2.9}$$

如果令 $t_i = x_i - x$ 为在区间$(x, x+1]$上第 i 个死亡者的生存时间，其中 $0 < t_i \leqslant 1$，则有

$$L_i = \frac{S(x+t_i)h(x+t_i)}{S(x)} = {}_{t_i}p_x(\mu_{x+t_i}) \tag{9.2.10}$$

将所有的死亡者的似然函数相乘，得到：

$$\prod_{i=1}^{d_x} {}_{t_i}p_x \cdot \mu_{x+t_i} \tag{9.2.11}$$

通常记为：

$$\prod_D {}_{t_i}p_x \cdot \mu_{x+t_i} \tag{9.2.12}$$

显然，对$(n_x - d_x)$个生存者，他们的似然函数为$(p_x)^{n_x-d_x} = (1-q_x)^{n_x-d_x}$，这样，总的似然函数为：

$$L = (1-q_x)^{n_x-d_x} \prod_D {}_{t_i}p_x \cdot \mu_{x+t_i} \tag{9.2.13}$$

为求解$\hat{q}_x$，需用 q_x 来表示${}_{t_i}p_x \cdot \mu_{x+t_i}$。在指数分布假设下，$\mu_{x+t}$为常数 μ，即 $\mu = -\ln p_x$。因此，

$${}_{t_i}p_x = (p_x)^{t_i} = e^{-\mu t_i} \tag{9.2.14}$$

似然函数可以写成：

$$L = (p_x)^{n_x-d_x} \cdot \prod_D [(p_x)^{t_i} \cdot \mu] = \mu^{d_x} \cdot \exp[-\mu(n_x - d_x) - \mu \cdot \sum_D t_i]$$

$$\ln L = d_x \cdot \ln\mu - \mu[(n_x - d_x) + \sum_D t_i]$$

令$\dfrac{d\ln L}{d\mu} = \dfrac{d_x}{\mu} - [(n_x - d_x) + \sum_D t_i] = 0$，解得：

$$\hat{\mu} = \frac{d_x}{(n_x - d_x) + \sum_D t_i},$$

所以，$\hat{\mu} = \dfrac{2}{98+0.2+0.6} = \dfrac{5}{247}$是 μ 的极大似然估计值，从而$\hat{q}_x = 1 - e^{-\hat{\mu}} = 1 - e^{-\frac{5}{247}} = 0.020039$。 ■

9.2.3 当删失和截断同时存在时的极大似然估计

在大多数情况下，右删失和左截断是同时存在的。例如，保险责任包括免赔和赔偿限额的保单理赔额；又或者，在特定观察期内，对某个观察对象的观测会分很多种不同情况：从初期开始观测、从期中开始观测；对象生存到观察期末、对象在观察期内死亡①。对各种情况下的似然函数，需要分情况来考虑：

1. 如果观测值 x_i 没有被截断或删失，则似然函数值为 $f(x_i)$；

2. 如果观测值 x_i 是在 d_i 被截断的，则似然函数值为 $f(x_i)/(1-F(d_i))$；

3. 如果观测值 x_i 是在 d_i 被截断且被平移的，则似然函数值为 $f(x_i+d_i)/(1-F(d_i))$；

4. 如果观测值 x_i 是在 u_i 被删失的，则似然函数值为 $1-F(u_i)$；

5. 如果观测值 x_i 是在 d_i 被截断且在 u_i 删失的，则似然函数值为 $\dfrac{1-F(u_i)}{1-F(d_i)}$。

注意到 $S(x)=1-F(x)$，$f(x)=h(x)S(x)$，因此对于没有被平移的数据，可以用一个公式来表示似然函数：

$$L=\prod_{i=1}^{n}\frac{S(x_i)[h(x_i)]^{\delta_i}}{S(d_i)},\ \delta_i=\begin{cases}1, & \text{未删失}\\0, & \text{在 } x_i \text{ 删失}\end{cases} \tag{9.2.15}$$

其中，若观测值 x_i 是在 u_i 被删失，则记 $x_i=u_i$。

【例 9-16】 考察表 9-7 所示的 6 个接受人工心脏的病人的样本。其中 4 人在 2006 年 12 月 31 日之前死亡（见表），观察期为日历年 2006 年。假设人工移植心脏病人的存活时间服从指数分布，由所给样本数据估计指数分布参数。

表 9-7 接受人工心脏的病人的生存情况

病人	移植时间	死亡时间
1	2005.1.1	2006.4.1
2	2005.4.1	2006.4.1
3	2005.7.1	—
4	2005.10.1	2006.7.1
5	2006.1.1	—
6	2006.4.1	2006.10.1

解： 由于观察期是日历年 2006 年，在此之前进行心脏移植手术的病人，在进入观察期时已经存活了一段时间，因此数据是被截断数据。而观察期结束后还存活的病人，其死亡时间未知，因此是删失数据。根据上述分析，这 6 个样本的截断点和真实值（删失点）为：$d_1=1$，$d_2=0.75$，$d_3=0.5$，$d_4=0.25$，$d_5=0$，$d_6=0$；$x_1=1.25$，$x_2=1$，$x_3=1.5$，$x_4=0.75$，

① 本章只考虑单风险环境下参数的估计，这时死亡是唯一的随机事件。在双风险环境中，死亡和退出都是随机事件。关于双风险环境下参数的估计，有兴趣的读者请参看李晓林、孙佳美主编的《生命表基础》，中国财政经济出版社 2006 年版。

$x_5=1$，$x_6=0.5$，其中 x_3、x_5 是删失点。则根据式（9.2.15），有

$$L=\prod_{i=1}^{6}\frac{e^{-x_i/\theta}\left(\frac{1}{\theta}\right)^{\delta_i}}{e^{-d_i/\theta}}$$

$$\ln L=\sum_{i=1}^{6}[-x_i/\theta+d_i/\theta+\delta_i\ln(1/\theta)]$$

令$\frac{d\ln L}{d\theta}=0$，得到：$\hat{\theta}=\sum_{i=1}^{6}(x_i-d_i)/\sum_{i=1}^{6}\delta_i=3.5/4=0.875$。

【例 9－17】 考察在区间$(x, x+1]$的 100 个观察对象，已知如下条件：

（1）有 70 个对象在 x 年年初开始被观测，那么截止到年底，观测到 16 个对象死亡

（2）有 30 个对象从$(x+0.6)$这个时点开始观察，30 个人中有 4 人会在年底前死亡。假设死亡力服从均匀分布（UDD）假设，用极大似然方法估计 q_x。

解： 用 T 来表示死亡时间。截止到年底，70 个在年初被观察的对象中有 16 个死亡，54 个生存。对于每一个对象来说死亡概率 $P[T\leqslant 1]=q$，生存概率 $P[T>1]=1-q$。对于 30 个在$(x+0.6)$才开始被观察的人来说，4 个会在年末前死亡，26 个生存。

注意，对于从$(x+0.6)$这个时点开始观察的 30 个人，它们在$(x, x+0.6]$上的信息是不可知的，所以它们在$(x+0.6)$处左截断。

死亡概率（在线性均匀分布假设下）：

$$P[T\leqslant 1\mid T>0.6]={}_{0.4}q_{0.6}=\frac{0.4q}{1-0.6q}$$

生存概率为：

$$1-{}_{0.4}q_{0.6}=1-\frac{0.4q}{1-0.6q}=\frac{1-q}{1-0.6q}$$

由此可以写出似然函数：

$$L(q)=q^{16}(1-q)^{54}\left(\frac{0.4q}{1-0.6q}\right)^{4}\left(\frac{1-q}{1-0.6q}\right)^{26}=(0.4)^4\frac{q^{20}(1-q)^{80}}{(1-0.6q)^{30}}$$

$$\ln L(q)=4\ln(0.4)+20\ln q+80\ln(1-q)-30\ln(1-0.6q)$$

$$\frac{d}{dq}\ln L(q)=\frac{20}{q}-\frac{80}{1-q}+\frac{30(0.6)}{1-0.6q}=0$$

解得$\hat{q}=0.238$。

【例 9－18】 已知观测到 20 例损失额（原始值，见表 9－8）。经验显示，损失额服从参数为 α 和 θ 的帕累托分布，且已知参数 $\theta=10\,000$，用极大似然法估计 α。

表 9－8　基本数据

损失额	观测数	免赔额	保单限额
750	3	200	∞
200	3	0	10 000
300	4	0	20 000
>10 000	6	0	10 000
400	4	300	∞

解： 帕累托分布的密度函数为 $f(x)=\dfrac{\alpha\theta^{\alpha}}{(\theta+x)^{\alpha+1}}$，所以可以写出似然函数：

$$L=\left[\frac{f(750)}{1-F(200)}\right]^{3}f(200)^{3}f(300)^{4}[1-F(10\ 000)]^{6}\left[\frac{f(400)}{1-F(300)}\right]^{4}$$

$$=\left[\frac{\alpha\ 10\ 200^{\alpha}}{10\ 750^{\alpha+1}}\right]^{3}\left[\frac{\alpha\ 10\ 000^{\alpha}}{10\ 200^{\alpha+1}}\right]^{3}\left[\frac{\alpha\ 10\ 000^{\alpha}}{10\ 300^{\alpha+1}}\right]^{4}\left[\frac{10\ 000^{\alpha}}{20\ 000^{\alpha}}\right]^{6}\left[\frac{\alpha\ 10\ 300^{\alpha}}{10\ 400^{\alpha+1}}\right]^{4}$$

$$\propto\alpha^{14}10\ 000^{13\alpha}10\ 750^{-3\alpha}20\ 000^{-6\alpha}10\ 400^{-4\alpha}$$

于是得到对数似然函数：

$$\begin{aligned}\ln L&=14\ \ln\alpha+13\alpha\ \ln(10\ 000)-3\alpha\ \ln(10\ 750)\\&\quad-6\alpha\ \ln(20\ 000)-4\alpha\ \ln(10\ 400)\\&=14\ \ln\alpha-4.5327\alpha\end{aligned}$$

对 α 求导令导数等于零，解得 $\hat{\alpha}=3.089$。 ■

§ 9.3 区间估计和方差

本节将计算极大似然估计值的区间估计和方差，以及估计值的某些函数的方差。在一般情况下，确定像极大似然估计这样复杂估计值的方差非常不容易，但还是有可能对其估计方差进行近似。其计算基础就是 Fisher 信息量。在这里我们只陈述定理的结论，不给出证明。其中 $L(\theta)$ 代表似然函数，$l(\theta)$ 代表对数似然函数，所有的结果都假设总体的分布是某个参数分布族的成员。

定理 9－1① 设 X_1，…，X_n 是独立同分布的，其概率密度函数（在离散情况下为概率函数）为 $f(x;\ \theta)$，θ 为参数，$f(x;\ \theta)$ 满足以下条件（对于离散情况下面的积分将换作求和）：

（1）参数空间 Ω 是一个开区间；

（2）集合 $A=\{x\,|\,f(x,\ \theta)>0\}$ 与参数 θ 的取值无关；

（3）对任意 $x\in A$，$f(x,\ \theta)$ 对 θ 三次可微，且二阶导数是 θ 的连续函数；

（4）积分 $\int f(x,\theta)dx$ 在积分号下二次可微，即 $\int\frac{\partial}{\partial\theta}f(x;\theta)dx=0$，$\int\frac{\partial^2}{\partial\theta^2}f(x;\theta)dx=0$；

（5）$-\infty<\int f(x;\theta)\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)dx<0$；

（6）存在函数 $H(x)$，使得对任意 $x\in A$，$\theta_0-c<\theta<\theta_0+c$，满足

① 定理的叙述和详细证明见 E. L. Lehmann, Theory of Point Estimation, 1983, John Wiley&Sons, pp415.

$\int H(x)f(x;\theta_0)dx < \infty$，并且 $\left|\frac{\partial^3}{\partial\theta^3}\ln f(x;\theta)\right| < H(x)$。

则有如下结论成立：

(a) 当 $n\to\infty$ 时，似然函数方程 $[L'(\theta)=0]$ 有解的概率趋近于 1；

(b) 当 $n\to\infty$ 时，极大似然估计值 $\hat{\theta}_n$ 的分布收敛到正态分布，均值为 θ，方差满足：$I(\theta)Var(\hat{\theta}_n)\to 1$，其中，

$$I(\theta) = -nE\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right] = -n\int f(x;\theta)\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)dx$$

$$= nE\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right] = n\int f(x;\theta)\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2 dx$$

定理的第二个结论表明，对于任意的 z，有

$$\lim_{n\to\infty}P\left(\frac{\hat{\theta}_n-\theta}{[I(\theta)]^{-1/2}}<z\right)=\Phi(z) \tag{9.3.1}$$

根据定理 9-1，$[I(\theta)]^{-1}$ 为 $Var(\hat{\theta}_n)$ 的一个有意义的估计，称 $I(\theta)$ 为 Fisher 信息量。由这个结果我们得到极大似然估计量是渐近无偏和相合的。信息量同时构成了 *Cramér - Rao* 下界，也就是说在一般条件下，任何无偏估计量的方差都大于由信息量的倒数给出的方差。因此，至少在渐近意义上，没有哪个无偏估计比极大似然估计更准确。(1) ~ (6) 给出的条件被称为"一般正则性条件"。这些条件通常是成立的，但是难以验证，所以只是假设这些条件都是成立的。① 这些条件将确保密度函数对参数比较光滑，并且密度函数比较正常。

上面的叙述假设样本由独立同分布的随机变量观测组成。更一般的结果用对数似然函数表示：

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}l(\theta)\right] = E\left[\left(\frac{\partial}{\partial\theta}l(\theta)\right)^2\right] \tag{9.3.2}$$

这里唯一的要求是每个观测的参数值相同。

如果有一个以上的参数，结论只是改为参数向量的极大似然估计服从渐近的多元正态分布。这个分布的协方差阵通过求一个 (r, s) 处元素，由下式给出的矩阵的逆得到：

$$I(\theta)_{rs} = -E\left[\frac{\partial^2}{\partial\theta_s\partial\theta_r}l(\theta)\right] = -nE\left[\frac{\partial^2}{\partial\theta_s\partial\theta_r}\ln f(X;\theta)\right]$$

$$= E\left[\frac{\partial}{\partial\theta_r}l(\theta)\frac{\partial}{\partial\theta_s}l(\theta)\right] = nE\left[\frac{\partial}{\partial\theta_r}\ln f(X;\theta)\frac{\partial}{\partial\theta_s}\ln f(X;\theta)\right]$$

① 如果总体分布是自然指数型的，这些条件成立，见陈希儒的《高等数理统计》，中国科学技术大学出版社 1999 年版，第 176 页。

其中第一和第三个表达式都是在一般情况下成立，第二和第四个表达式假设似然函数是 n 个相同密度函数的乘积。这个矩阵通常称为“信息阵”。为保证 $I^{-1}(\theta)$ 存在，通常要求信息矩阵 $I(\theta)$ 是正定的。

【例 9－19】 估计对数正态分布的极大似然估计量的协方差阵。然后计算例 9－1 数据的相应结果。

解： 似然函数和对数似然函数为：

$$L(u,\sigma)=\prod_{j=1}^{n}\frac{1}{x_j\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln x_j-u)^2}{2\sigma^2}\right]$$

$$l(u,\sigma)=\sum_{j=1}^{n}\left[-\ln x_j-\ln\sigma-\frac{1}{2}\ln(2\pi)-\frac{1}{2}\left(\frac{\ln x_j-u}{\sigma}\right)^2\right]$$

它们的一阶导数为：

$$\frac{\partial l}{\partial u}=\sum_{j=1}^{n}\frac{\ln x_j-u}{\sigma^2}\text{和}\frac{\partial l}{\partial\sigma}=-\frac{n}{\sigma}+\sum_{j=1}^{n}\frac{(\ln x_j-u)^2}{\sigma^3}$$

二阶导数为：

$$\frac{\partial^2 l}{\partial\mu^2}=-\frac{n}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial\sigma\partial\mu}=-2\sum_{j=1}^{n}\frac{\ln x_j-u}{\sigma^3}$$

$$\frac{\partial^2 l}{\partial\sigma^2}=\frac{n}{\sigma^2}-3\sum_{j=1}^{n}\frac{(\ln x_j-u)^2}{\sigma^4}$$

由于 $\ln X_j$ 服从均值为 u，标准差为 σ 的正态分布，因此，

$$E\left(\frac{\partial^2 l}{\partial u^2}\right)=-\frac{n}{\sigma^2},E\left(\frac{\partial^2 l}{\partial u\partial\sigma}\right)=0,E\left(\frac{\partial^2 l}{\partial\sigma^2}\right)=-\frac{2n}{\sigma^2}$$

改变符号后求逆可得到协方差阵的估计：

$$\begin{bmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{\sigma^2}{2n}\end{bmatrix}$$

对于对数正态分布，极大似然估计是下面两个方程的解：

$$\sum_{j=1}^{n}\frac{\ln x_j-u}{\sigma^2}=0\text{ 且 }-\frac{n}{\sigma}+\sum_{j=1}^{n}\frac{(\ln x_j-u)^2}{\sigma^3}=0$$

由第一个等式得到 $\hat{u}=(1/n)\sum_{j=1}^{n}\ln x_j$，由第二个等式有 $\hat{\sigma}^2=(1/n)\sum_{j=1}^{n}(\ln x_j-\hat{u})^2$。 ■

对于例 9－1 的数据，这些值分别为 $\hat{u}=9.127$ 和 $\hat{\sigma}^2=1.563$。协方差矩阵需要参数的真值，但能够做到的最好方法是使用估计值代替真值。于是得到参数 μ，σ 的极大似然估计值的方差估计值：

$$\widehat{Var}(\hat{u},\hat{\sigma})=\begin{bmatrix}0.07815 & 0\\ 0 & 0.03908\end{bmatrix}$$

主对角线外的 0 说明这两个参数估计渐近不相关。在对数正态分布这个特殊情形下，这个结论对任何容量的样本都是成立的。因此可以构造一个置信度为 95% 的参数真值的置信区间，在估计值两边的 1.96 个标准差内有

$$u: 9.127 \pm 1.96(0.07815)^{1/2} = 9.127 \pm 0.548$$

$$\sigma: 1.25 \pm 1.96\ (0.03908)^{1/2} = 1.25 \pm 0.3871$$

为了得到信息阵需要计算导数和期望值，这通常并不容易。避免这个问题的一个方法是不计算期望值，直接使用观测数据。这个结果称为“已观测信息量”。

【例 9－20】　使用已观测信息量估计前例的协方差阵。

解： 将观测值代入二阶导数方程得到：

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} = -\frac{20}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial \sigma \partial \mu} = -2\sum_{j=1}^{n}\frac{\ln x_j - u}{\sigma^3} = -2\,\frac{182.54 - 20u}{\sigma^3}$$

$$\frac{\partial^2 l}{\partial \sigma^2} = \frac{n}{\sigma^2} - 3\sum_{j=1}^{n}\frac{(\ln x_j - u)^2}{\sigma^4} = \frac{20}{\sigma^2} - 3\,\frac{1\,697.2487 - 365.08u + 20u^2}{\sigma^4}$$

带入参数的估计值得到观测信息的负值项：

$$\frac{\partial^2 l}{\partial \mu^2} = -12.7959,\ \frac{\partial^2 l}{\partial \sigma \partial \mu} = 0,\ \frac{\partial^2 l}{\partial \sigma^2} = -25.5259$$

改变符号求逆得到了和前面相同的结果，这是对数正态分布的特点，对于其他分布并不一定成立。■

信息矩阵提供了一种评价参数的极大似然估计质量的方法。但是，人们更关心作为参数函数的某些量的估计。比如，我们希望以帕累托分布的均值作为总体均值的估计。也就是说，以 $\hat{\alpha}\hat{\theta}/(\hat{\alpha}-1)$ 作为总体均值的一个估计，其中采用了极大似然估计值。计算这个随机变量的均值和方差是很困难的，因为其中的两个随机变量的分布已经相当的复杂。对于参数函数的估计，常用方法为 delta 方法。定理 9－2 给出了该方法的描述。

定理 9－2　令 $X_n = (X_{1n}, \cdots, X_{kn})^T$ 表示一个容量为 n 的 k 维多元随机变量样本。假设 X 服从渐近正态分布，均值为 θ，协方差阵为 Σ/n，其中 θ 和 Σ 都不依赖于 n，g 是一个完全可微的 k 元函数，$G_n = g(X_{1n}, \cdots, X_{kn})$，则 G_n 也服从均值为 $g(\theta)$、方差为 $(\partial g)^T\Sigma(\partial g)/n$ 的渐近正态分布。其中，∂g 为一阶导数向量 $\partial g = (\partial g/\partial\theta_1,\ \cdots,\ \partial g/\partial\theta_k)^T$，而且取值于原随机变量参数的真值 θ。

定理 9－2 的陈述比较难解释。X 是估计量，g 为待估参数的函数。对于单参数模型，这个定理的陈述如下：令 $\hat{\theta}$ 为 θ 的估计量，服从均值为 θ、方差为 σ^2/n 的渐近正态分布。则 $g(\hat{\theta})$ 也服从渐近正态分布，均值为 $g(\theta)$，

渐近方差为：

$$[g'(\theta)](\sigma^2/n)[g'(\theta)]=[g'(\theta)]^2\sigma^2/n \tag{9.3.3}$$

【例 9－21】 使用 delta 方法给出指数分布超过 4 000 的概率值的极大似然估计的方差，然后计算例 9－1 数据的结果。

解： 指数分布参数 θ 的极大似然估计是样本均值。要估计的量为：

$$p=P(X>4\ 000)=\exp(-4\ 000/\theta)$$

其极大似然估计为：

$$\hat{p}=\exp(-4\ 000/\hat{\theta})=\exp(-4\ 000/\bar{x})$$

计算这个估计量的均值和方差并不容易，但是已知 $Var(\bar{X})=Var(X)/n=\theta^2/n$，进一步地，有

$$g(\theta)=e^{-4\ 000/\theta},g'(\theta)=4\ 000\theta^{-2}e^{-4\ 000/\theta}$$

因此，由 delta 方法，有

$$Var(\hat{p})=\frac{(4\ 000\theta^{-2}e^{-4\ 000/\theta})^2\theta^2}{n}=\frac{16\ 000\ 000\theta^{-2}e^{-8\ 000/\theta}}{n}$$

对于例 9－1 的数据，有

$$\bar{x}=22\ 626.5$$

$$\hat{p}=\exp(-4\ 000/22\ 626.5)=0.8380$$

$$Var(\hat{p})=\frac{16\ 000\ 000\ (22\ 626.5)^{-2}e^{-8\ 000/22\ 626.5}}{n}=0.001097$$

因此，p 的 95% 置信区间为 $0.838\pm1.96\sqrt{0.001097}$，即（0.773，0.903）。

■

【例 9－22】 利用例 9－1 的数据构造对数正态总体均值的 95% 置信区间。将这个结果与由传统方法样本均值得到的置信区间作比较。

解： 由例 9.3.2 得到 $\hat{u}=9.127$ 和 $\hat{\sigma}=1.2502$，两个估计量的协方差矩阵为：

$$\frac{\Sigma}{n}=\begin{bmatrix}0.07815 & 0\\ 0 & 0.03908\end{bmatrix}$$

函数 $g(u,\sigma)=\exp(\mu+\frac{1}{2}\sigma^2)$，偏导数为：

$$\frac{\partial g}{\partial u}=\exp\left(u+\frac{1}{2}\sigma^2\right),\frac{\partial g}{\partial\sigma}=\sigma\exp\left(u+\frac{1}{2}\sigma^2\right)$$

这些量的估计值分别为 20 100.5 和 25 129.65。由 delta 方法得到下面的估计：

$$\begin{aligned}Var[g(\hat{u},\hat{\sigma})]&=[20\ 100.5\quad 25\ 129.6]\begin{bmatrix}0.07815 & 0\\ 0 & 0.03908\end{bmatrix}\begin{bmatrix}20\ 100.5\\ 25\ 129.6\end{bmatrix}\\&=56\ 253\ 771.05\end{aligned}$$

置信区间为 20 100.5 ± 14 700.5，即(5 400，34 801)。■

按照传统方法，总体均值的置信区间为样本均值的邻域 $\bar{x} \pm 1.96s/\sqrt{n}$，这里 s^2 为样本方差。对于例 9－1 的数据，这个区间为 22 626.5 ± 20 594.1。这是一个更大的区间，这并不奇怪，因为我们知道（对于对数正态分布）极大似然估计量是渐近的一致最小方差无偏估计。

【例 9－23】 已知累积危险率函数 $H(5)$ 的 95% 置信区间为(0.283，1.267)，使用 delta 方法求 $S(5)$ 的 95% 的置信区间。

解： 该置信区间的中点为 0.775，即 $\hat{H} = 0.775$。由

$$\left(0.775 - 1.96\sqrt{Var(\hat{H}(5))}, 0.775 + 1.96\sqrt{Var(\hat{H}(5))}\right) = (0.283, 1.267)$$

知 $Var(\hat{H}(5))$ 的估计值为 0.063。因为 $S(t) = \exp(-H(t))$，$\dfrac{dS(t)}{dH(t)} = e^{-H(t)}$，由 delta 方法，

$$Var(\hat{S}(t)) = (e^{-\hat{H}(t)})^2 Var(\hat{H}(t))$$

代入　　$\hat{H}(5) = 0.775, Var(\hat{H}(5)) = 0.063$，得：

$$Var(\hat{S}(5)) = (e^{-0.775})^2 0.063 = 0.0134$$

$S(5)$ 的估计值为 $e^{-0.775} = 0.4607$，置信区间为 $0.4607 \pm 1.96\sqrt{0.0134} = (0.23, 0.69)$。

§ 9.4　多变量的参数模型

9.4.1　引言

前面所讨论的参数模型都是针对单变量的，在精算模型中所关心的随机变量也可能受到其他变量的影响，或者受到被保险人的风险特征变量的影响。本节将考虑这两种情况下的多变量参数模型。

当随机变量依赖于其他变量的取值，可建立相互依赖变量模型。例如联合生命年金或者寿险的情形，保险赔付的时间依赖于第一个或者第二个个体的死亡。因为这些个体通常是相关的（典型的如夫妇），死亡时间相互依赖。

按照年金领取人分类，我们所熟知的年金保险可以分为个人年金和联合生存年金。**个人年金**是指以一个被保险人的生存作为给付条件的年金保险；**联合生存年金**是指以两个或两个以上的被保险人的生命作为给付年金的条件的年金。联合生存年金又可以分为共同生存年金保险和最后生存者年金保险。以两个被保险人为例，前者是指如果联合投保人中有一人死亡，

年金给付即行停止；后者是指直到所有被保险人都死亡，年金给付才停止。

【例9-24】 一对夫妇购买了共同生存年金保险。购买该保险时，丈夫64岁，妻子63岁。假设丈夫的生存函数为 $S_X(t)=\frac{36-t}{36},0\leqslant t\leqslant 36$；妻子的生存函数为 $S_Y(t)=\frac{37-t}{37}$，$0\leqslant t\leqslant 37$。假设夫妇二人的死亡时间独立（在大多数联合生存模型中，两个个体的生存时间一般是相互影响的，此处的独立性假设只是为了建模方便，更符合现实的情况将在9.4.2中讨论），那么在第10年末时，此共同生存年金保险的给付仍未终止的概率是多少呢？

解：显然，只有当两个人在第10年末时都生存，共同生存年金保险的给付才不会终止。设共同生存年金保险的未来给付时间为 T，即求 $P(T\geqslant 10)=S_T(10)$，因为夫妇二人的死亡时间独立，可得：

$$P(T\geqslant 10)=S_T(10)=S_X(10)\cdot S_Y(10)=0.73\times 0.72=0.527$$ ■

当随机变量的分布可能会依赖于被保险人的特定属性时，考虑建立包含伴随变量的多变量模型是比较合适的。例如，考虑 x 岁的人在 $x+t$ 岁时仍生存的概率 $S(t)=P(T>t)$，其中 T 表示 x 岁的人的未来寿命。显然不同年龄的人在 t 年后仍然生存的概率是不同的。因此，在建立精算生存模型时，考虑年龄对生存函数的影响是十分必要的。这时，生存函数就变为 $S(t;x)=P(T>t)$。除了年龄之外，投保人的职业、地域、吸烟与否、血压、身高、体重等都会对未来寿命造成影响，这些因素都可以作为伴随变量被加入到生存模型中。又如考虑机动车辆的年索赔数，这个随机变量的分布可能与机动车的驾驶里程、驾驶地区和主要驾驶员的状况如年龄、性别、婚姻状况和驾驶记录有关。

假设我们考虑的目标变量 X，受 m 个伴随变量 z_1，z_2，…，z_m 的影响，令 $\mathbf{Z}=(z_1, z_2, \cdots, z_m)'$，$S(t\mid\mathbf{Z})$ 代表加入伴随变量 $\mathbf{Z}$ 之后的生存函数，$h(t\mid\mathbf{Z})$ 表示相应的危险率函数，X 的密度函数为 $f(t\mid\mathbf{Z})$。我们可以得到如下关系：

$$h(t\mid\mathbf{Z})=-\frac{d}{dt}\ln S(t\mid\mathbf{Z}) \tag{9.4.1}$$

$$S(t\mid\mathbf{Z})=\exp[-\int_0^t h(w\mid\mathbf{Z})dw \tag{9.4.2}$$

$$f(t\mid\mathbf{Z})=S(t\mid\mathbf{Z})h(t\mid\mathbf{Z}) \tag{9.4.3}$$

那么我们只要能够得到 $S(t\mid\mathbf{Z})$ 和 $h(t\mid\mathbf{Z})$ 的数学表达式，就可以通过样本数据，计算似然函数 L，对 L 求最大值来估计生存模型中的参数了。又根据式(9.4.1)和(9.4.2)，我们实际上只需要知道 $h(t\mid\mathbf{Z})$ 的形式，就可以进行参数估计了。但是，由于伴随变量以及其他许多因素的影响，$S(t$

$|\mathbf{Z})$ 和 $h(t|\mathbf{Z})$ 的形式有时非常复杂。因此，常常需要对 $h(t|\mathbf{Z})$ 或 $S(t|\mathbf{Z})$ 的形式进行假设。

【例 9－25】 假设汽车保险的事故次数与驾驶员的年龄和性别有关，请对该保险的事故次数建立合适的模型。

解： 可以有三种不同的建模方式：

（1）对每一个性别和年龄的组合分别建立模型。这需要搜集每一个组合的事故次数数据。该种方法的缺点是繁琐，待估参数个数多，有可能存在数据量不足的问题。

（2）建立统一的参数模型。譬如，假设给定驾驶员的事故次数服从参数为 λ 的泊松分布，而泊松参数又与年龄（x）和性别（$y=1$ 为男性，$y=0$ 为女性）有关，如：

$$\lambda=(\alpha_0+\alpha_1x+\alpha_2x^2)\beta^y \text{ 或者 } \lambda=e^{\alpha x+\beta y}$$

此方法的特点是参数较少，可以用通常的方法进行估计，比较适合上述的车险事故次数。

（3）从一个密度函数、分布函数或危险率函数模型出发，用年龄和性别对这个函数进行修正。

$$S(n|x;y)=[S_0(n)]^{(\alpha_0+\alpha_1x+\alpha_2x^2)\beta^y}$$

显然后两个模型较第一个模型更优，因为某些年龄和性别的组合可能没有观察值或观察值很少。此时，第一个模型就无法应用，而后两个模型仍然可以应用这些信息，并且待估参数的个数较少。当给定的个体不满足任何明显的分布模型时，第三个模型将是不错的选择。对于汽车保险的例子，泊松分布是一个合理的选择，因此第二个模型可能是最好的方法。

9.4.2　二元变量模型

正如在例 9－23 中介绍的那样，同时考虑相互依赖的二元变量构造的模型就称为“二元模型”。二元模型的理论基础就是概率论中联合密度函数和边缘密度函数。

显然，如果我们所考虑的两个变量之间独立，那么二元变量的密度（分布）函数就等于这两个变量的边缘密度（分布）函数的乘积。因此，只要已知边缘分布，构造二元分布就变得非常简单。

但是，在实际问题中，我们所考虑的两个变量往往是相关的。例如，沿用之前联合生存年金的例子，夫妇二人的生存时间通常并不是独立的，丈夫的死亡可能使妻子的寿命缩短。此时，将边缘分布相乘直接得到二元分布就不合理了，就需要将两变量之间的相关性引入模型。引入相关性的方法有很多种，这里只介绍在精算实务中应用比较广泛的耦合（Copula）分布。

耦合分布是采用一个耦合函数进行构造的，这个函数本身必须为单位正方形上的一个正规的二元分布函数，其边缘分布为均匀分布。用 $F_X(x)$ 和 $F_Y(y)$ 表示两个边缘分布函数，$C(u, v)$ 表示耦合函数。由这三个函数构造的二元分布为：

$$F_{X,Y}(x,y) = C[F_X(x), F_Y(y)] \tag{9.4.4}$$

耦合函数比较简单的形式为 $C(u,v) = uv$，由此构造的二元分布函数为 $F_{X,Y}(x,y) = F_X(x)F_Y(y)$，相互独立变量的联合分布就是这个表达式。但是正如前面所讨论的那样，现实问题中我们所研究的变量之间通常是相关的，所以这种耦合函数的形式并不常用。

比较常用的耦合函数形式为 Frank **耦合函数**，其具体形式如下（其中 $\log_\alpha$ 表示以 α 为底的对数）：

$$C(u,v) = \log_\alpha\left[1 + \frac{(\alpha^u - 1)(\alpha^v - 1)}{\alpha - 1}\right] \tag{9.4.5}$$

参数 α 用来控制两个变量之间的关联性。α 的值小于 1 表示正向关联性，大于 1 表示反向关联，等于 1 表示两变量之间独立。则式（9.4.4）就变成如下形式：

$$F_{X,Y}(x,y) = \log_\alpha\left[1 + \frac{(\alpha^{F_X(x)} - 1)(\alpha^{F_Y(y)} - 1)}{\alpha - 1}\right] \tag{9.4.6}$$

令 $\beta = \ln\alpha$，则式（9.4.6）也可以写成：

$$F_{X,Y}(x, y) = \frac{1}{\beta}\ln\left[1 + \frac{(e^{\beta F_X(x)} - 1)(e^{\beta F_Y(y)} - 1)}{e^\beta - 1}\right] \tag{9.4.7}$$

这样，如果我们给定了单个变量的边缘分布函数 $F_X(x)$，$F_Y(y)$ 以及参数 α，就可以得到二元变量模型了。

【例 9-26】 设一对夫妇购买了最后生存者年金保险（即只要有一个被保险人生存，年金的给付就不会终止），购买保险时丈夫 60 岁，妻子 50 岁。刚出生的个体的剩余寿命服从 De Moivre 定律，极限年龄 $\varpi = 100$。假设两人的剩余寿命间的关系由 Frank 耦合函数表示，$\alpha = e^{0.5}$，求此共同生存年金保险在 10 年内停止给付的概率。

解： 设丈夫的剩余寿命随机变量为 X，妻子的剩余寿命变量为 Y。由 De Moivre 定律，可求得边缘分布为：

$$F_X(x) = \frac{x}{40}, F_Y(y) = \frac{y}{50}$$

已知 $\alpha = e^{0.5}$，则 $\beta = 0.5$，代入公式（9.4.7），可计算二元分布函数为：

$$\begin{aligned}F_{X,Y}(x,y) &= \frac{1}{\beta}\ln\left[1 + \frac{(e^{\beta F_X(x)} - 1)(e^{\beta F_Y(y)} - 1)}{e^\beta - 1}\right] \\ &= \frac{1}{0.5}\ln\left[1 + \frac{(e^{0.5\frac{x}{40}} - 1)(e^{0.5\frac{y}{50}} - 1)}{e^{0.5} - 1}\right]\end{aligned}$$

此共同生存年金保险在 10 年内停止给付的概率为：

$$F_{X,Y}(10,10)=\frac{1}{0.5}\ln\left[1+\frac{(e^{0.5\frac{10}{40}}-1)(e^{0.5\frac{10}{50}}-1)}{e^{0.5}-1}\right]=0.0427$$ ■

【例 9－27】 假设索赔数据变量 X 和相应的直接理赔费用 Y 通常具有很强的正相关性，假设它们的边缘分布均为帕累托分布，使用 Frank 耦合确定联合分布的模型。

解： 设 X 的边缘分布参数为 β，θ，则 $F_X(x)=1-\left(\frac{\theta}{x+\theta}\right)^{\beta}=1-\left(1+\frac{x}{\theta}\right)^{-\beta}$；Y 的边缘分布参数为 γ，τ，则 $F_Y(y)=1-\left(\frac{\tau}{y+\tau}\right)^{\gamma}=1-\left(1+\frac{y}{\tau}\right)^{-\gamma}$，根据式（9.4.3），可得二元分布函数：

$$F_{X,Y}(x,y)=\log_{\alpha}\left\{1+\frac{[\alpha^{1-(1+x/\theta)^{-\beta}}-1][\alpha^{1-(1+y/\tau)^{-\gamma}}-1]}{\alpha-1}\right\}$$

对 x 和 y 求偏导，得到联合密度函数：

$$f_{X,Y}(x,y)=\frac{(\alpha-1)\frac{\beta\gamma}{\theta\tau}\alpha^{2-(1+x/\theta)^{-\beta}-(1+y/\tau)^{-\gamma}}\times(1+x/\theta)^{-\beta-1}(1+y/\tau)^{-\gamma-1}\ln\alpha}{\{\alpha-1+[\alpha^{1-(1+x/\theta)^{-\beta}}-1][\alpha^{1-(1+y/\tau)^{-\gamma}}-1]\}^{2}}$$

对 X 的密度函数构造极大似然函数，利用样本数据估计 X 边缘分布的参数 β，θ，同理可以估计 γ，τ。将参数估计值 $\hat{\beta}$，$\hat{\theta}$，$\hat{\gamma}$，$\hat{\tau}$ 代入联合密度函数的极大似然函数中，则可以根据样本数据估计参数 α，并进而得到联合密度函数。 ■

9.4.3　和模型

和模型是一种含伴随变量的参数模型。它将总危险率 $h(t|\mathbf{Z})$ 表示成基本危险率 $h(t)$ 和由伴随变量引起的附加危险率的和的形式。假设第 j 个伴随变量在时刻 t 处引起的附加危险率可以表示为 $h(t|z_j)=h_j(t)\cdot g_j(z_j)$，则可得到：

$$h(t|\mathbf{Z})=h_0(t)+\sum_{j=1}^{s}h_j(t)g_j(z_j) \tag{9.4.8}$$

特别地，有 $g_j(z_j)=z_j$（对于所有的 j）。并令 $h_j(t)=a_j$，则总危险率模型可简化为：

$$h(t|\mathbf{Z})=h_0(t)+\sum_{j=1}^{s}a_jz_j \tag{9.4.9}$$

当基本危险率为一常数 a_0 时，令 $z_0=1$，就可以得到总危险率的最终简化形式了。

$$h(t|\mathbf{Z})=\sum_{j=0}^{s}a_jz_j=a'z=a_0+a_1z_1+a_2z_2\cdots+a_sz_s \tag{9.4.10}$$

其中，$a'=(a_0,a_1,\cdots,a_s)$，$\mathbf{Z}=(z_0,z_1,\cdots,z_s)'$，且 $z_0=1$。

显然，根据总危险率的最终简化形式，总危险率是一个与变量 X 无关的量。在生存模型中，如果生存函数服从指数分布，在常值死力假设下，总危险率模型是伴随变量 z 的线性形式，这时总危险率模型的最终简化形式又被称为线性指数模型。

有了 $h(t\mid \mathbf{Z})$ 的表达式，我们就可以利用极大似然估计法来估计模型的参数。考虑生存模型，设观察对象 i 在时间 d_i 开始进入观察，并在时间 t_i 离开观察，离开观察的方式有退保或死亡。Z_i 是指第 i 个对象的所有伴随变量 z_{i1}，z_{i2}，…，z_{is}的观测值向量。那么根据第二节的讨论，第 i 个对象对应的似然函数为：

$$L_i=\begin{cases}\dfrac{f(t_i\mid Z_i)}{1-F(d_i\mid Z_i)}, & \text{在时刻 } t_i \text{ 死亡}\\[2ex] \dfrac{1-F(t_i\mid Z)}{1-F(d_i\mid Z_i)}, & \text{在时刻 } t_i \text{ 退保}\end{cases}$$

由式（9.4.3）知，似然函数可以表示为：

$$L_i=\frac{S(t_i\mid Z_i)[h(t_i\mid Z_i)]^{\delta_i}}{S(d_i\mid Z_i)},\quad \delta_i=\begin{cases}1, & \text{第 } i \text{ 个对象在 } t_i \text{ 时刻死亡}\\ 0, & \text{第 } i \text{ 个对象在 } t_i \text{ 时刻退保}\end{cases}$$

总似然函数为：

$$L=\prod_{i=1}^{n}[h(t_i\mid Z_i)]^{\delta_i}\frac{S(t_i\mid Z_i)}{S(d_i\mid Z_i)} \tag{9.4.11}$$

其中，$\dfrac{S(t_i\mid Z_i)}{S(d_i\mid Z_i)}=\exp\left[-\int_{d_i}^{t_i}h(\omega\mid Z_i)d\omega\right]=\exp\left[-\int_{d_i}^{t_i}a'Z_i d\omega\right]=\exp[-a'Z_i(t_i-d_i)]$

则总似然函数为：

$$L=\prod_{i=1}^{n}[a'Z_i]^{\delta_i}\exp[-a'Z_i(t_i-d_i)]$$

对其求对数：

$$\ln L=\sum_{i=1}^{n}\delta_i\ln(a'Z_i)-\sum_{i=1}^{n}a'Z_i(t_i-r_i)$$

根据极大似然估计原理$\dfrac{\partial\ln L}{\partial a_j}=0$，得到：

$$\sum_{i=1}^{n}\frac{\delta_i z_{ij}}{a'Z_i}-\sum_{i=1}^{n}z_{ij}(t_i-d_i)=0 \tag{9.4.12}$$

这里 $a'Z_i=(a_0,a_1,a_2,\cdots,a_s)(z_{i0},z_{i1},z_{i2},\cdots,z_{is})'=a_0+a_1z_{i1}+a_2z_{i2}+\cdots+a_sz_{is}$。

这里我们需要注意，由于示性函数 δ_i 的存在，能够得到的方程的个数≤待估参数的个数 $s+1$，所以有时需要使用数值分析的方法得到参数的估计值$\hat{a}_0$，$\hat{a}_1,\cdots,\hat{a}_s$。

【例 9－28】 对一线性指数模型，第 i 个观察对象的危险率函数为 $h(t\mid z_i)=a_1z_{i1}+a_2z_{i2}$，其中 z_{ij}为第 i 个对象的第 j 个伴随变量的观测值，a_1，

a_2 为参数。所有的对象均由 $t=0$ 开始观察。已知表 9－9 中包含 4 人组成的样本的数据，求 a_1，a_2 的极大似然估计值。

表 9－9　　例 9－27 的观察数据

观察对象 i	离开时间	离开方式	伴随变量	
			z_{i1}	z_{i2}
1	1	死亡	2	1
2	2	死亡	1	1
3	3	终止	4	4
4	5	终止	8	8

解： 因为所有的对象均由 $t=0$ 开始观察，所以对于所有的 i，$r_i=0$。又根据各个对象退出观察的方式，知道 $\delta_1=\delta_2=1$，$\delta_3=\delta_4=0$。

$$a'Z_1=a_1z_{11}+a_2z_{12}=2a_1+a_2$$
$$a'Z_2=a_1z_{21}+a_2z_{22}=a_1+a_2$$
$$a'Z_3=a_1z_{31}+a_2z_{32}=4a_1+4a_2$$
$$a'Z_4=a_1z_{41}+a_2z_{42}=8a_1+8a_2$$

将已知数据代入 $\sum_{i=1}^{n}\frac{\delta_i z_{ij}}{a'Z_i}-\sum_{i=1}^{n}z_{ij}(t_i-r_i)=0$

对于 $j=1$，

$$\frac{2}{2a_1+a_2}+\frac{1}{a_1+a_2}-[2\times1+1\times2+4\times3+8\times5]=0 \tag{9.4.13}$$

对于 $j=2$，

$$\frac{1}{2a_1+a_2}+\frac{1}{a_1+a_2}-[1\times1+1\times2+4\times3+8\times5]=0 \tag{9.4.14}$$

整理式（9.4.13）和（9.4.14）：

$$\begin{cases}2a_1+a_2=1\\ a_1+a_2=\dfrac{1}{54}\end{cases}$$

解得：$a_1=\frac{53}{54}$，$a_2=-\frac{52}{54}$

第 i 个观察对象的危险率函数为：

$$h(t\mid Z_i)=\frac{53}{54}z_{i1}-\frac{52}{54}z_{i2}$$

生存函数为：

$$S(t;Z_i)=\exp\left[-\int_0^t h(\varpi\mid Z_i)d\varpi\right]=\exp\left[\frac{52}{54}Z_{i1}t-\frac{53}{54}Z_{i2}t\right]$$ ■

9.4.4 积模型

积模型就是将总危险率由基本危险率 $h(t)$ 与由伴随变量产生的附加危险率相乘得到。其中最常用的模型是 Cox 模型（也称**比例风险模型**）。已知基本危险率为 $h_0(t)$，伴随变量为 z_0，$z_1,\cdots,z_s$，Cox 模型就是：

$$h(t\mid Z)=h_0(t)c(a_1z_1+a_2z_2+\cdots+a_sz_s)=h_0(t)c(a'Z) \tag{9.4.15}$$

其中 $c(y)$ 为自变量的定义域为正数的任意函数，$a'=(a_1,a_2,\cdots,a_s)$，$Z=(z_1,z_2,\cdots,z_s)'$。

一般情况下，Cox 模型指 $c(y)=e^y$ 的情况。则 Cox 模型的基本结构是：

$$h(t\mid Z)=h_0(t)\exp\left(\sum_{j=1}^{s}a_jz_j\right)=h_0(t)e^{a'Z} \tag{9.4.16}$$

按照 Cox 模型的形式，易知构建 cox 模型需要满足如下条件：

1. 比例风险假定（PH 假定，proportional hazards）。任两个个体危险率函数之比不随时间改变。即

$$\frac{h_i(t)}{h_j(t)}=\exp\{a'(Z_i-Z_j)\},\ i,\ j=1,\cdots,n,\ h_i(t)=h(t\mid Z_i)$$

2. 对数线性假定。任一个体 i 的危险率函数的对数与伴随变量 Z_i 呈线性关系。即

$$\ln h_i(t)-\ln h_0(t)=a'Z_i$$

在生存模型中，在常值死力假设下，$h_0(t)=e^{a_0}$，则可以得到最终简化模型：

$$h(t\mid Z)=\exp\left(\sum_{j=0}^{s}a_jz_j\right)=e^{a'Z} \tag{9.4.17}$$

其中 $a'=(a_0,\ a_1,\cdots,a_s)$，$Z=(z_0,\ z_1,\cdots,z_s)'$，且 $z_0=1$。若生存模型服从指数分布，则 cox 模型下的总危险率模型又被称为“对数线性指数模型”。

在常值死力假设下由式（9.4.11）

$$L=\prod_{i=1}^{n}[h(t_i\mid Z_i)]^{\delta_i}\frac{S(t_i\mid Z_i)}{S(d_i\mid Z_i)}$$

其中，$h(t_i\mid Z_i)=e^{a'Z_i}$

$$\begin{aligned}\frac{S(t_i;Z_i)}{S(d_i;Z_i)}&=\exp\left[-\int_{d_i}^{t_i}h(\omega\mid Z_i)d\omega\right]=\exp\left[-\int_{d_i}^{t_i}e^{a'Z_i}d\omega\right]\\&=\exp[-e^{a'Z_i}(t_i-d_i)]\end{aligned}$$

则总似然函数

$$L=\prod_{i=1}^{n}[e^{a'Z_i}]^{\delta_i}\exp[-e^{a'Z_i}(t_i-d_i)]$$

对其求对数得：

$$\ln L = \sum_{i=1}^{n} \delta_i(a'Z_i) - \sum_{i=1}^{n} e^{a'Z_i}(t_i - d_i)$$

令$\frac{\partial \ln L}{\partial a_j}=0$，得到：

$$\sum_{i=1}^{n} \delta_i z_{ij} - \sum_{i=1}^{n} z_{ij}(t_i - d_i)e^{a'Z_i} = 0 \tag{9.4.18}$$

求解关于 a_j 的方程组，我们就可以得到参数的估计值$\hat{a}_0$，$\hat{a}_1$，…，$\hat{a}_s$ 了。

【例 9－29】 对一个肝癌患者群体，其生存函数与被观察对象的起始年龄、是否酗酒及自起始年龄开始的时间有关，假设对数线性指数生存模型为适宜模型。设 z_1 为被观察对象的起始年龄，z_2 为表示是否酗酒的示性函数，$z_2=1$ 表示酗酒；对一个由 6 个病人组成的样本，由 $t=0$ 开始观察。表 9－10 给出了 z_1 和 z_2 的值，以及死亡观察对象退出观察的时间和方式，确定可以求解$\hat{a}_0$，$\hat{a}_1$，$\hat{a}_2$ 的方程。

表 9－10　　例 9－28 中的观察数据

i（病人）	z_1	z_2	死亡时间	观察终止时的生存时间
1	51	1	5.0	—
2	52	0	—	4.5
3	53	0	—	3.0
4	55	1	8.0	—
5	60	1	—	7.0
6	61	0	6.0	—

解： 因为所有的对象均由 $t=0$ 开始观察，所以对于所有的 i，$d_i=0$。根据各个对象退出观察的方式，知道 $\delta_1=\delta_4=\delta_6=1$，$\delta_2=\delta_3=\delta_5=0$。

$$a'Z_1=a_0+51a_1+a_2,\qquad a'Z_2=a_0+52a_1$$
$$a'Z_3=a_0+53a_1,\qquad a'Z_4=a_0+55a_1+a_2$$
$$a'Z_5=a_0+60a_1+a_2,\qquad a'Z_6=a_0+61a_1$$

对于 $j=0$，$\sum_{i=1}^{n} \delta_i z_{ij} - \sum_{i=1}^{n} z_{ij}(t_i - d_i)e^{a'Z_i} = 0$ 可化为：

$$3-[5e^{a_0+51a_1+a_2}+4.5e^{a_0+52a_1}+3e^{a_0+53a_1}+8e^{a_0+55a_1+a_2}+7e^{a_0+60a_1+a_2}+6e^{a_0+61a_1}]=0$$

对于 $j=1$，得到：

$$51+55+61-(51\times 5e^{a_0+51a_1+a_2}+52\times 4.5e^{a_0+52a_1}+53\times 3e^{a_0+53a_1}$$
$$+55\times 8e^{a_0+55a_1+a_2}+60\times 7e^{a_0+60a_1+a_2}+61\times 6e^{a_0+61a_1})=0$$

对于 $j=2$，得到：

$$2-(5e^{a_0+51a_1+a_2}+8e^{a_0+55a_1+a_2}+7e^{a_0+60a_1+a_2})=0$$

■

【例 9－30】 假设机动车辆险的索赔额依赖于车辆的价值和车辆的使用性质（营运或非营运），假设基本危险率服从指数分布，分别建立比例风险模型。

解：设 z_1 代表车辆的价值，显然 z_1 的取值为非负；z_2 代表车辆的使用性质；$z_2=1$ 代表营运车，$z_2=0$ 代表非营运的车。则任意一个车辆的危险率函数形式为：

$$h(x \mid \mathbf{Z}) = h_0(x)e^{\beta_1 z_1 + \beta_2 z_2}$$

考虑两个价值均为 z_1 的营运车和非营运车，危险率函数之间存在如下关系：

$$h_{营运}(x) = h_0(x)e^{\beta_1 z_1 + \beta_2} = h_{非营运}(x)e^{\beta_2}$$

根据生存函数和危险率函数之间的关系，可以得到两个车辆的生存函数之间的关系：

$$S_{营运}(x) = \exp\left[-\int_0^x h_{营运}(y)dy\right] = \exp\left[-\int_0^x h_{非营运}(y)e^{\beta_2}dy\right]$$
$$= [S_{非营运}(x)]^{\exp(\beta_2)}$$

基本危险率可以使用参数模型或者样本数据来估计，然后使用极大似然估计方法估计 β_1，β_2。

为了构造极大似然函数，我们需要求出生存函数和密度函数。令 $c_j = \exp(\beta^T z)$ 表示第 j 个观测的 cox 乘数，则 $S_j(x) = S_0(x)^{c_j}$，其中 $S_0(x)$ 是基本危险率对应的生存函数。密度函数为：

$$f_j(x) = -\frac{\partial S_j(x)}{\partial x} = -c_j S_0(x)^{c_j-1}(S_0(x))' = c_j S_0(x)^{c_j-1} f_0(x)$$

对于基本危险率服从指数分布的情况，有

$$S_j(x) = [e^{-x/\theta}]^{c_j} = e^{-c_j x/\theta}, \quad f_j(x) - \left(\frac{c_j}{\theta}\right)e^{-c_j x/\theta}$$ ■

9.4.5 广义线性模型

和模型和积模型要求生存函数之间具有特殊的关系。从精算的角度来说，这可能不是最合理的，因为很难解释危险率函数乘以一个常数的含义，或对生存函数求幂的意义。考虑将变量与我们关心的量例如期望值直接建立关系，可能更加有意义。常用的方法是回归分析法。

回归分析是研究一个变量关于另一个（些）变量的依赖关系的计算方法和理论，其目的在于用后者的已知或设定值，去估计（预测）前者的（总体）均值。较为广泛使用的回归模型是古典线性回归模型，但是由于精算问题中所采用数据的特殊性，无法满足古典线性回归模型的假设，因此近些年来，**广义线性回归模型**（generalized linear models，GLM）在精算中迅速得到应用。

1. 古典线性回归模型。在介绍广义线性回归模型之前，先来回顾一下

古典线性回归模型的知识。模型形式为：

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n+\varepsilon$$

其中 Y 表示被解释变量（因变量），x 表示解释变量（自变量），ε 是随机误差，β_i 为未知参数，称为回归系数。在古典线性回归模型假定：

第一，解释变量 x_1，x_2，…，x_n 是非随机变量，线性无关；

第二，独立正态性假定：$\varepsilon_i|x\sim N(0,\sigma^2)$，$Cov(\varepsilon_i,\varepsilon_j|x)=0$，$i\neq j$。

根据模型的假定，可知古典线性回归模型具有如下特征：

（1）$E(Y)=x'\beta=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n$；

（2）x、Y 都是取连续值的变量，如农作物的产量，人的身高体重之类；

（3）Y 的分布为正态，或接近正态之分布。

尽管古典回归分析在经济、金融中得到了大量应用，但是由于保险数据的特点，古典回归分析在用于处理保险数据时存在一定的缺陷，主要体现为：

首先，古典线性回归模型要求随机误差项，进而因变量服从正态分布。但在保险数据中，这一假设常常不能保证，尤其是非寿险的损失分布，通常具有不对称、定义域非负并且尾部较厚的特点，此时使用正态分布的假设就不太合适了。

其次，古典线性回归模型假设误差项同方差，从而因变量的方差也为常数。但是，精算问题中这一假设常常会遭到破坏，例如当赔付额越大时，其误差也应该越大才合理，即方差应该随均值的改变而发生变化，所以要求同方差的假定不现实。

再次，损失事件发生的概率是 0 和 1 之间的数，若将此变量作为因变量，简单地将概率表示为解释变量的线性组合是不合理的，因为线性组合的取值范围并不局限于（0，1）。

最后，古典线性回归模型要求用解释变量的线性表达式来解释被解释变量，但有时这种解释关系可能是乘法关系。

2. 指数分布族。广义线性模型将被解释变量服从正态分布的假定推广到服从指数分布族。因此我们首先来了解一下指数型分布族，其概率密度函数可以表示为：

$$f(y_i;\theta_i,\phi)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right\} \tag{9.4.19}$$

其中 $a(\cdot),b(\cdot),c(\cdot)$ 为已知函数，它们对所有的观测值形式相同。$a(\cdot)$的形式通常为 ϕ/ϖ，其中 ϖ 是先验权重；函数 $b(\cdot)$称为“累积函数”，θ_i 称为“自然参数”，其取值与均值有关；ϕ_i 称为“离散参数”，与方差有关。

根据密度函数的性质，易知有如下等式存在：

$$\int_{-\infty}^{\infty} f(y_i;\theta_i,\phi)dy_i = 1 = \int_{-\infty}^{\infty} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}dy_i$$

设 $\mu_i = E(Y_i)$，则对上式求 θ_i 的一阶二阶三阶导数，得到如下结果：

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{9.4.20}$$

$$Var(Y_i) = a(\phi)b''(\theta_i) \tag{9.4.21}$$

$$E[(Y_i - \mu_i)^3] = a^2(\phi)b'''(\theta_i) \tag{9.4.22}$$

由上述结果可以得到如下性质和结论：

（1）指数型分布族的均值 μ_i 由自然参数 θ_i 唯一确定。若函数 $b'(\cdot)$ 存在逆函数，则可得 $\theta_i = b'^{-1}(\mu_i)$。

（2）指数型分布的方差是两个函数的乘积，一个是关于自然参数 θ_i 的函数 $b''(\theta_i)$，如(1)所述，也是均值 μ_i 的函数，被称为“方差函数”；另一个是关于离散参数 ϕ 的函数 $a(\phi)$，与自然参数和均值无关。如果将方差函数表示成均值的函数，记为 $V(\mu_i)$，则有 $Var(Y_i) = \frac{\phi V(\mu_i)}{\varpi_i}$，其中 $V(\mu_i)$ 为方差函数；ϕ 为离散参数，将方差进行了缩放；ω_i 表示第 i 个观察值的权重。

指数分布族包含了许多常见分布，例如正态分布、二项分布等等。表 9－11列出了几种常见的指数分布族的参数形式。

表 9－11　　　　指数分布族中的常见分布

分　布	$a(\phi)$	$b(\theta)$	$c(y,\ \phi)$
正态分布	ϕ/ω	$\theta^2/2$	$-0.5\ [\omega y^2/\phi + \ln\ (2\pi\phi/\omega)]$
泊松分布	ϕ/ω	$\exp\ (\theta)$	$-\ln\ (y!)$
二项分布（试验 m 次）	ϕ/ω	$m\cdot\ln\ [1+\exp\ (\theta)]$	$\ln\binom{m}{y}$
伽玛分布	ϕ/ω	$-\ln(-\theta)$	$\frac{\omega}{\phi}\ln\left(\frac{\omega y}{\phi}\right) - \ln(y) - \ln\left[\Gamma\left(\frac{\omega}{\phi}\right)\right]$
逆高斯分布	ϕ/ω	$-\sqrt{-2\theta}$	$-0.5\ [\ln\ (2\pi\phi y^3\omega + \omega/\ (\phi y)]$

3. 广义线性模型。广义线性模型从以下几个方面对古典线性模型进行了推广：

（1）$E(Y) = \mu = g^{-1}(x'\beta)$（注意古典线性回归模型为 $E(Y) = \mu = x'\beta$），g 为一严格单调、充分光滑的函数，称为“联结函数”。联结函数的反函数称为“均值函数”。记 $\eta \doteq g(\mu) = x'\beta$。

（2）x，Y 可取连续或离散值，且在应用上更多见的情况为离散值，如 $\{0，1\}$，$\{0，1，\cdots，\infty\}$ 等。

（3）Y 的分布属于指数型分布族，正态分布是其一特例。$a(\theta)$的选取与 y 的分布有关。

由此看出，广义线性模型是由三部分构成的：随机成分、系统成分、联结函数。

- **随机成分**，因变量 Y 或误差项的概率分布。与古典线性回归模型中假设因变量的每个观察值相互独立服从正态分布不同，广义线性模型假设因变量的每个观察值相互独立且服从指数型分布族中的一个分布。
- **系统成分**，即解释变量的线性组合，表示为 $\eta_i = x'\beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$。
- **联结函数**，联结函数建立了随机成分与系统成分之间的关系。与古典线性回归模型中将因变量 Y 的期望值直接等于解释变量的线性组合不同，广义线性模型将因变量 Y 的期望值表示成解释变量线性组合的函数变换，这个变换就利用到了联结函数。虽然不同的观测值可以有不同的联结函数，但是实际应用中通常对所有的观测值使用相同的联结函数。

表 9－12　　常见的联结函数

联结函数名称	$g(\mu_i)=\eta_i$	$\mu_i=g^{-1}(\eta_i)$
恒等	μ_i	η_i
对数	$\ln(\mu_i)$	$\exp(\eta_i)$
倒数	μ_i^{-1}	η_i^{-1}
Logit	$\ln[\mu_i/(1-\mu_i)]$	$1/[1+\exp(-\eta_i)]$

显然，古典线性回归模型就是广义线性模型联结函数取恒等形式，因变量来自于正态分布的特例。

在广义线性模型中，因变量来自指数分布族，所以方差函数可以表示成均值的函数，这一特性非常适合保险数据，这也是广义线性模型在费率厘定等精算问题中应用广泛的原因之一。

一般地，广义线性模型中因变量分布形式的确定主要依靠保险数据的先验信息，其主要根据就是因变量的方差与均值的关系。例如，如果因变量的方差为常数，则可以选择正态分布；如果因变量的方差等于均值，则可以选择泊松分布；如果因变量的方差等于均值的平方，则可以选择伽玛分布；如果因变量的方差等于其均值的三次方，可以选择逆高斯分布。

根据保险数据自身的特点，我们在估计不同的因变量时，常使用一些典型的广义线性模型：

（1）在估计索赔次数或索赔频率时，典型的广义线性模型是泊松乘法模型，即使用对数联结函数和泊松分布的误差项。在估计索赔次数时，通常用风险单位数加权。在例 9－31 中，我们将看到如何利用泊松乘法模型

来估计索赔次数。

（2）在估计索赔强度时，一般使用伽玛乘法模型，即使用对数联结函数和伽玛分布的误差项。

（3）在估计续保率和新业务转换率时，常使用 Logistic 模型，即采用 Logit 联结函数和二项分布的误差项。该模型在因变量的取值很小时，可以用泊松乘法广义线性模型近似。

与古典线性回归模型类似，广义线性模型中的参数估计也可采用极大似然估计的方法。下面用一个例子来说明广义线性模型的求解及在精算中的应用。

【例 9－31】 设汽车保险的风险分类系统包括两个变量：行驶区域和驾驶员性别。行驶区域的取值为城市和乡村两个水平，驾驶员性别的取值为男和女两个水平。假设实际观察的保单组索赔次数如表 9－13 所示，分别建立古典线性回归模型和广义线性模型将索赔次数表示成行驶区域和性别的线性组合。

表 9－13　索赔次数的观察值

性别	城市	乡村
男	4 000	2 500
女	2 000	1 000

解： 设索赔次数为 Y，男性为 x_1、女性为 x_2、城市为 x_3、乡村为 x_4，并且这四个变量的取值只能为 0 或 1，那么对于男性驾驶员，$x_1=1$、$x_2=0$，对于女性驾驶员 $x_1=0$，$x_2=1$，对于行驶在城市的车辆 $x_3=1$，$x_4=0$，对于行驶在乡村的车辆 $x_3=0$，$x_4=1$。

（1）古典线性回归模型。因为 $x_1+x_2+x_3+x_4=2$，解释变量之间存在多重共线性，所以去掉一个解释变量 x_4，建立古典线性回归模型：

$$Y=\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$$

按照最小二乘法得到参数估计值为：

$$\beta_1=2\ 625,\ \beta_2=875,\ \beta_3=1\ 250$$

则模型为 $Y=2\ 625x_1+875x_2+1\ 250x_3$。

根据参数估计值，可得索赔次数的拟合值如表 9－14 所示。

表 9－14　索赔次数的拟合值：古典线性回归模型

性别	城市	乡村
男	3 875	2 625
女	2 125	875

（2）广义线性模型。将风险分为男性城市、男性乡村、女性城市、女性乡村 4 个类别。设索赔次数为 Y，男性为 x_1、女性为 x_2、城市为 x_3、乡村为 x_4，并且四个解释变量的取值只能为 0 或 1。为了避免多重共线性，剔除 x_4，则各个类别的期望索赔次数可以表示为：

$$\mu_i = g^{-1}(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})$$

索赔次数服从泊松分布，则有 $f(y;\mu) = \frac{\mu^y}{y!}e^{-\mu}$，似然函数为：

$$L = \prod_{i=1}^{n} f(y_i,\mu_i) = \prod_{i=1}^{n}\left[\frac{\mu_i^{y_i}}{y_i!}e^{-\mu_i}\right]$$

对似然函数求对数得：

$$l = \sum_{i=1}^{n} \ln f(y_i,\mu_i) \sum_{i=1}^{n} [y_i \ln(\mu_i) - \ln(y_i!) - \mu_i] \tag{9.4.23}$$

选用对数联结函数，则有 $\mu_i = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})$，即

$$\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\\ \mu_4\end{bmatrix} = \begin{bmatrix}\exp(\beta_1+\beta_3)\\ \exp(\beta_1)\\ \exp(\beta_2+\beta_3)\\ \exp(\beta_2)\end{bmatrix} \tag{9.4.24}$$

将观察值 y_i 和式（9.4.24）代入式（9.4.23）中，得到似然函数的对数为：

$$l = [4\ 000(\beta_1+\beta_3) - \ln 4\ 000! - \exp(\beta_1+\beta_3)] + [2\ 500\beta_1 - \ln 2\ 500! - \exp(\beta_1)] + [2\ 000(\beta_2+\beta_3) - \ln 2\ 000! - \exp(\beta_2+\beta_3)] + [1\ 000\beta_2 - \ln 1\ 000! - \exp(\beta_2)]$$

求 $\frac{\partial l}{\partial \beta_1} = \frac{\partial l}{\partial \beta_2} = \frac{\partial l}{\partial \beta_3} = 0$，则有

$$\frac{\partial l}{\partial \beta_1} = 4\ 000 - \exp(\beta_1+\beta_3) + 2\ 500 - \exp(\beta_1) = 0$$

$$\frac{\partial l}{\partial \beta_2} = 2\ 000 - \exp(\beta_2+\beta_3) + 1\ 000 - \exp(\beta_2) = 0$$

$$\frac{\partial l}{\partial \beta_3} = 4\ 000 - \exp(\beta_1+\beta_3) + 2\ 000 - \exp(\beta_2+\beta_3) = 0$$

解上述方程组可得参数估计值为 $\beta_1 = 1.7810$，$\beta_2 = 7.0078$，$\beta_3 = 0.539$。根据参数估计值，可得索赔次数的拟合值如表 9－15 所示。

表 9－15　索赔次数的拟合值：泊松分布假设和对数联结函数

性别	城市	乡村
男	4 105.2	2 394.7
女	1 894.7	1 105.2

■

习　题

1. 已知索赔额分布服从伽玛分布，随机的 10 个索赔额样本：1 500、6 000、3 500、3 800、1 800、5 500、4 800、4 200、3 900、3 000，试用矩估计法估计参数 α 和 θ。

2. 假设索赔额分布为帕累托分布，随机 20 个索赔额样本为：27、82、115、126、155、161、243、294、340、384、457、680、855、877、974、1 193、1 340、1 884、2 558、15 743。计算矩估计参数。

3. X 是密度为 $f(x) = px^{p-1}(0 < x < 1)$ 的连续随机变量，一组随机样本的三次观测为 0.2、0.3、0.5。找出极大似然估计的 p。

4. 运用中位数估计法对习题 2 中的数据进行参数估计，假设分布为指数分布。

5. 一组分组数据具有如下性质：

$$(0,5]-n_1=2,(5,10]-n_2=2,(10,25]-n_3=2,(25,\infty]-n_4=2$$

运用极大似然估计方法估计下面分布的参数：（1）参数为 θ 的指数分布；（2）（0，θ）上的均匀分布。

6. 以下是对于限额为 20 的保单的 10 次赔付额：3、5、6、8、9、13、16、20、20、20（三个 20 均为赔偿限额，所以这三次的损失额都大于 20）。假设损失额服从［0，θ］的均匀分布。运用矩估计方法估计 θ。

7. 对于 10 个有免赔额 5 的保单赔付额：4、5、6、6、9、14、16、18、22、25。假设损失额服从以下几种分布、运用矩估计参数 θ：（1）［0，θ］上的均匀分布；（2）参数为 θ 的指数分布；（3）帕累托分布，参数为 θ 和 $\alpha=2$。

8. 下面是一组赔偿限额为 50 的某险种保单赔付额：3、4、8、10、12、18、22、35、50、50、50、50。如果损失额 X 服从指数分布，运用极大似然估计方法估计参数。

9. 对于第 8 题，假设 X 服从（0，θ）上的均匀分布，运用极大似然估计方法估计参数。

10. 假设某类保单的免赔额为 5，随机抽取了 8 张保单的理赔额如下：3、4、8、10、12、18、22、35。假设损失额服从以下两种分布，运用极大似然估计方法估计参数：（1）指数分布；（2）（0，θ）上的均匀分布。

11. 假设 X 服从帕累托分布，参数 $\theta=10$，随机抽取了 8 个样本：3、4、8、10、12、18、22、35。求参数 α 的极大似然估计以及 $P(X\leqslant 10)$ 的极大似然估计。

12. 已知某类保单的免赔额 $d=2$，赔偿限额为 $u=16$，随机抽取了 8 次的赔付额观测值为：1、2、6、8、10、14、14、14。假设初始损失额分布为参数为 θ 的指数分布，求 θ 的极大似然估计。

13. 一组免赔额为 5 的保单赔付样本为：6、7、7、9、11、17、21、34。假设初始损失额服从指数分布，求参数 θ 的极大似然估计。

14. 已知随机变量 X 服从韦伯分布，密度函数为 $f(x) = \dfrac{\tau\,(x/\theta)^{\tau} e^{-(x/\theta)^{\tau}}}{x}$，随机抽取 8 个样本：3、4、8、10、12、18、22、35。已知参数 $\tau=0.374$，

求 θ 的极大似然估计以及 P（$X \leqslant 10$）的极大似然估计。

15. 假设 X 是服从参数 θ 的指数分布，12 个随机样本为：7、12、15、19、26、27、29、29、30、33、38、53。求 θ 的极大似然估计和 $I(\theta)$。

16. 运用第 15 题的数据，求 $Var(X)$ 的估计的方差。

17. 利用第 15 题的数据，计算 $P[X > 10]$ 的置信度为 95% 的置信区间。

18. 随机变量 X 的 12 个样本点为：7、12、15、19、26、27、29、29、30、33、38、53。假设 X 服从指数分布，求参数估计的方差的极大似然估计。

19. 求第 18 题中的方差的极大似然估计，求方差估计值的 95% 的置信区间。

第十章 参数模型的检验和选择

学习目标

- □ 了解选择最优模型的方法
- □ 学会运用 $p-p$ 图、Q－Q 图和平均剩余生命图等图形来直观选择合适的分布
- □ 掌握 χ^2 拟合优度检验、K－S 检验、Anderson－Darling 检验和似然比检验等选择比较分布

§10.1 引 言

在建模过程的最后阶段，我们需要从众多的备选模型中选择一个“最优”模型。这里的“最优”定义取决于具体的应用，依赖于对所研究的问题的了解。实际上我们要从模型的使用条件、拟合程度和使用的局限性等各个方面来对众多备选模型进行检验和筛选。

我们知道，图像是显示差异的最简单、直观的方法。因此在对比数据和模型时，图像的比较是最方便快捷的途径。但是，图像并不总是可靠的，特别是在备选模型的分布函数与数据经验分布函数比较接近时，很难从图像上分辨出细微的差距。此时，统计学假设检验的方法可以更有逻辑地显示出不同模型间的拟合程度差异。构造的检验统计量往往是对模型的分布函数和经验分布函数之间接近程度的一个度量。最后，图像比较和假设检验的结果可以引导我们进行模型的合理选择，确定“最优”的拟合模型。但是，无论选择哪个模型，都是对实际情况的一种近似。所有的模型都存在一定的问题，并不存在绝对最优的模型。不过，我们选择的“最优”模型还是可以在一定程度上使用，得到的最终结论也有意义。

一般来说，对模型的筛选将经历如下的过程：

- 被选模型与实际数据图形上的直观比较和筛选；
- 用统计学方法对模型分布函数与经验分布函数进行检验（如 χ^2 拟合优度检验、K－S 检验、Anderson－Darling 检验等）；
- 由一定的标准进行模型选择（常用主观判断法和评分法）。

本章中，我们将讨论模型评估的几种方法，进而总结出一个特定的建模策略，可以适用于大多数情形下的模型选择。需要注意的是，每种方法都有自身的优势和劣势，选择不同的方法，可能得到不同的模型。因此，

模型筛选在作为一门知识的同时也是一种艺术。

§10.2 模型的直观选择

10.2.1 数据与模型的表示

模型是对数据分布规律的展示。模型通常可以通过密度函数和分布函数来表示，或者是由其决定的函数，例如平均剩余寿命函数或者带上限的期望值函数。数据通常用经验分布函数或者直方图来表示，其中经验分布函数用于表示个体数据，而直方图常用于表示分组数据。对于个体完整数据，这些经验图形是很容易得到的。但是对于分组数据、截断数据和删失数据来说，会有一定的难度。对于离散模型，很少存在删失、截断以及分组的情况，数据可以由每个观测点处的相对频率和累积频率来表示。

本章中只讨论在同一点截断或删失数据。假设数据集的截断点是 t，则经验分布的起始点也是 t。为了和经验值进行比较，我们使用的模型必须是截断的。因此，截断后的模型表示为：

$$F^{*}(x)=\begin{cases}0, & x<d\\ \dfrac{F(x)-F(d)}{1-F(d)}, & x\geqslant d\end{cases}\qquad f^{*}(x)=\begin{cases}0, & x<d\\ \dfrac{f(x)}{1-F(d)}, & x\geqslant d\end{cases}$$

其中 $F(x)$、$f(x)$ 表示没有截断的模型。

本章中，如果分布函数或密度函数的下标为样本容量，则表示它是经验模型下的分布函数或密度函数，如 $F_n(x)$；如果没有标记或者用（ * ）标注，则表示为估计的参数模型。

10.2.2 密度函数与分布函数的图像比较

对模型拟合程度最直接的检验方法是做图。我们一般选用经验分布图（卵形图）、直方图、核密度图等与备选模型的分布函数或密度函数图进行比较。当模型与样本的分布图像比较接近时，我们认为可以使用该函数拟合样本数据。如果差异较大，超出了可以接受的范围，则认为不能使用该函数进行拟合。下面用几个例子说明图像比较法的具体应用。

【例 10-1】 由 10 只试验老鼠组成的样本，其死亡时间（以天为单位）为：3、4、5、7、7、8、10、10、10、12。假定适合的生存模型为指数分布，试用极大似然估计指数分布参数，并用图像对比法进行模型的筛选。

解：设指数分布的分布函数为：$F(x)=1-\exp(-x/\theta)$，用极大似然估计可以得到指数分布的参数为：$\hat{\theta}=7.6$，$\hat{F}(x)=1-\exp(-x/7.6)$。将

样本经验分布函数与估计函数画在同一个图中（见图 10－1），由图像可以看出，用指数分布来拟合小鼠生存函数并不合适。在 x 较小时，拟合值大于样本值。当 x 较大时，拟合值小于样本值。■

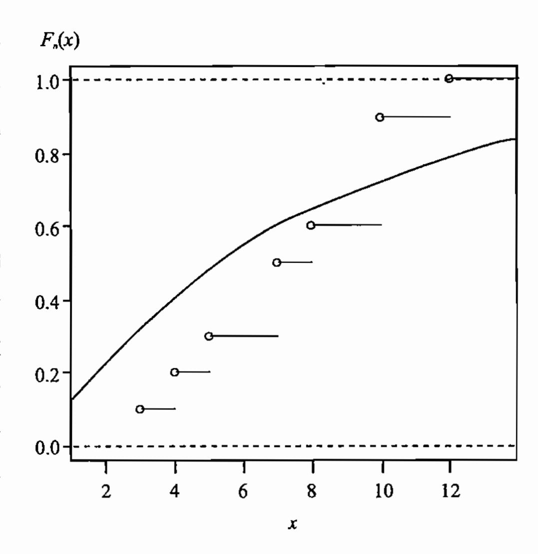


图 10－1　小鼠死亡时间的经验分布函数图像

【例 10－2】 考虑如下的 20 个车险赔付数据（见表 10－1），假设数据在 50 处被截断，用做图法检验可否用指数分布对该数据样本进行拟合？指数分布的参数由样本数据估计得出。

表 10－1　　车险赔付数据

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1 193	1 340	1 884	2 558	3 476

解： 若数据在 50 处截断，截断后的样本数据为 19 个，若使用指数分布对该数据样本进行拟合，由式（9.2.6）知，似然函数为 $\prod_{j=2}^{20}\dfrac{f(x_j,\theta)}{1-F(50)}$，用极大似然估计可以得出该指数分布的参数为：$\hat{\theta}=802.32$，则分布函数为：

$$F(x)=1-\exp(-0.001246x+0.0623)$$

将样本点的经验分布以及估计的指数分布图像编程绘制到一张图像上（见图 10－2），并对比样本经验分布与指数分布的差异大小。

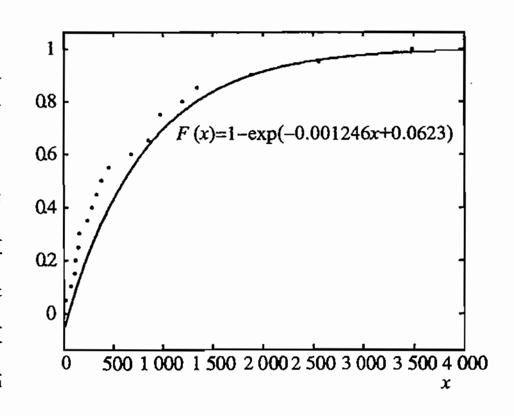


图 10－2　样本经验分布与指数分布的差异

观察图 10－2 可知，指数分布的分布函数图像基本可以显示数据的大小与变化，但是在 x 较小时拟合效果并不理想，样本点明显高于分布函数图像，这表明在 x 取较小值时模型低估了样本。■

当模型的分布函数和经验分布函数很接近时，很难从图像上分辨出细

微的差别。可以直接画出两个函数差值的图像。也就是说，如果 $F_n(x)$ 和 $F^*(x)$ 分别表示经验分布函数和由模型得到的分布函数，画出 $D(x)=F_n(x)-F^*(x)$ 的图像即可。

图 10－3 为例 10－1 中的 $D(x)$ 图像。

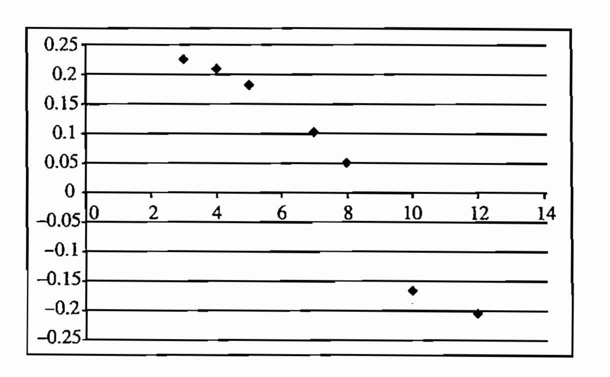


图 10－3　例 10－1 中数据的 $D(x)$ 图像

由图 10－3 可见 $F_n(x)$ 和 $F^*(x)$ 的差距在 ±0.25 之间，差距较大，因此并不适合用指数分布来拟合样本数据，与例子结论一致。

10.2.3　*p*－*p* 图和 Q－Q 图比较

p－*p* 图，也称“概率图”（probability plot），根据变量的经验分布与指定分布的累积分布函数之间的关系所绘制的图形。通过 *p*－*p* 图可以检验数据是否符合指定的分布。其步骤如下：首先将观测值排序 $x_1\leqslant\cdots\leqslant x_n$，再对每个值构造坐标（$F_n(x_j)$，$F^*(x_j)$），最后将每个坐标对应的点画在（$F_n(x)$，$F_n^*(x)$）的平面上。当数据符合指定分布时，*p*－*p* 图中各点近似呈一条 45°直线。但是，在这种情况下，必须对经验分布函数的定义有所修改。因为可以证明，$F_n(x_j)$ 的期望值为 $j/(n+1)$，进而经验分布函数在该点的值也应当是这个值而非通常的取值 j/n。对两个相同的观测值可以直接标为两个点（它们有相同的“*y*”坐标但“*x*”坐标不同），也可以取“*x*”坐标的平均值只画一个点。

如果 *p*－*p* 图中各点不呈直线，但有一定规律，则可以对变量数据进行转换，使转换后的数据更接近指定分布。

【例 10－3】　由 10 只试验老鼠组成的样本，其死亡时间（以天为单位）为：3、4、5、7、7、8、10、10、10、12。画出指数分布的 *p*－*p* 图。

解： 我们用 x_1，…，x_{10} 来表示上面各值。以观测值 $x_{(2)}=4$ 为例，经验分布值为 $F_{10}(4)=2/11=0.1818$，另一个坐标的值是：

$$F^*(4)=1-\exp(-4/7.6)=0.4092$$

这就得到 *p*－*p* 图中的一个点（0.1818，0.4092）。类似地，可以得到所有的点，其图形见图 10－4。

从图 10－5 可以看出，指数分布不适合描述该数据的分布。　■

Q－Q 图同样可以用于检验数据的分布。所不同的是，Q－Q 图是用样本数据的经验分位数与所指定分布的分位数之间的关系曲线来进行检验的。图 10－5 是例 10－1 的数据的 Q－Q 图。

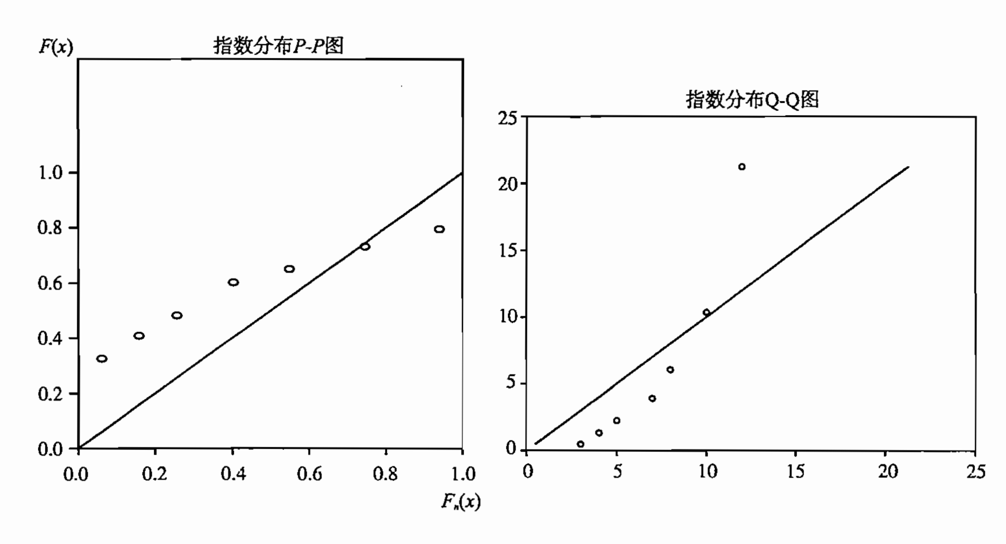


图 10－4　例 10－1 的数据 $p-p$ 图　　　图 10－5　例 10－1 数据的 Q－Q 图

当分析 $p-p$ 图和 Q－Q 图时，最好不要用严格的标准去衡量这些数据是否在一条直线上，通常只要看这些点是否近似在一条直线上即可。另外，当判断概率图上的点是否近似在一条直线上时，对样本点中两端的点（即太大和太小的样本点）可以不用关注，除非这些点偏离直线特别远（因为两端的少数样本点通常会和直线有些偏离）。但是，当有一个样本点偏离直线特别远，而其他样本点又基本近似在直线上时，偏离直线的那个样本点则视为离群点，不用考虑。

10.2.4　平均剩余寿命函数图

平均剩余寿命函数考虑的是数据在尾部的情况，其定义为：

$$e(d)=E[X-d \mid X>d] \tag{10.2.1}$$

如果平均剩余寿命函数随 d 递增，那么在变量取值较大处的期望结果会很大，因此概率向右移，说明其尾部相比那些平均剩余寿命函数递减或增速较慢的模型更厚。反之，如果平均剩余寿命函数随 d 递减，说明 X 的分布是轻尾分布。这里通过样本平均剩余寿命函数图$(d,\hat{e}(d))$观察样本数据的尾部特征。使用经验估计 $\hat{e}(d)$ 来代替 $e(d)$，有

$$\hat{e}(d)=\frac{\sum_{i=1}^{n}\max(X_i-d,0)}{\sum_{i=1}^{n}I_{\{X_i>d\}}} \tag{10.2.2}$$

如果平均剩余寿命函数图呈现上升的趋势，说明样本的损失分布是一个明显的厚尾分布；而如果呈现下降的趋势则是轻尾分布；指数分布的平均超额函数图近似为一条水平的直线。

【例 10－4】　表 10－2 为某地区男性经验生命表（1990—1993 年）的

部分数据，请用平均剩余寿命函数说明指数分布是否合适描述该数据。

表 10－2 某地区男性经验生命表及平均期望寿命表

年龄	${}_nM_x$	${}_nq_x$	l_x	${}_nd_x$	${}_nl_x$	T_x	$\mathring{e}_x$
<1	0.03451	0.03347	100 000	3 347	96 988	7 075 361	70.8
1～4	0.00185	0.00736	96 653	711	384 904	6 978 374	72.2
5～9	0.00058	0.0029	95 941	279	479 011	6 593 470	68.7
10～14	0.00045	0.00227	95 663	217	477 772	6 114 459	63.9
15～19	0.00083	0.00415	95 446	396	476 239	5 636 687	59.1
20～24	0.00105	0.00525	95 050	499	474 001	5 160 448	54.3
25～29	0.00109	0.00546	94 550	516	471 463	4 686 447	49.6
30～34	0.00127	0.00633	94 035	595	468 686	4 214 984	44.8
35～39	0.0017	0.00847	93 440	792	465 220	3 746 297	40.1
40～44	0.00249	0.01235	92 648	1 144	460 380	3 281 077	35.4
45～49	0.00392	0.01939	91 504	1 774	453 084	2 820 697	30.8
50～54	0.00629	0.03098	89 730	2 780	441 698	2 367 613	26.4
55～59	0.01007	0.04911	86 950	4 270	424 074	1 925 914	22.1
60～64	0.01635	0.07856	82 680	6 495	397 161	1 501 840	18.2
65～69	0.02669	0.12512	76 184	9 532	357 092	1 104 679	14.5
70～74	0.04475	0.20125	66 652	13 414	299 726	747 587	11.2
75～79	0.07518	0.31641	53 238	16 845	224 078	447 861	8.4
80～84	0.12427	0.47406	36 393	17 252	138 833	223 784	6.1
85～89	0.19587	0.65743	19 140	12 583	64 243	84 951	4.4
90～94	0.29391	0.78096	6 557	5 121	17 423	20 708	3.2
95～99	0.42003	0.85022	1 436	1 221	2 907	3 285	2.3
100+	0.56948	1	215	215	378	378	1.8

如果可以假设生存函数就是这样的一条连接各个点的直线，则可以计算平均剩余寿命函数，可以用给定年龄的曲线下方的面积除以该年龄的生存函数计算。图 10－6 为平均剩余寿命函数的图形。出生后不久的微量上升说明 1990 年婴儿的死亡率较高，出生一年后的生存个体的期望剩余寿命会多 5 年。随后的平均剩余寿命平稳递减，这符合我们预期的衰老模式。

比较平均剩余寿命图可以发现，指数分布并不适合描述该模型。指数分布图像是水平线的，而由观察可知，剩余寿命图像刚开始是上凸，然后

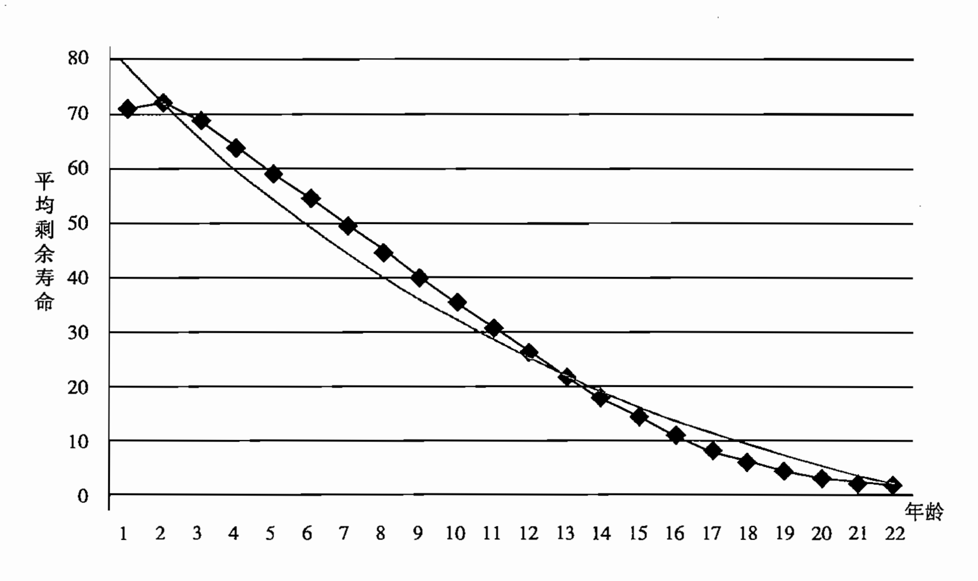


图 10－6　平均剩余寿命图像

下凸，与指数分布图像不符。 ■

§10.3　分布的拟合优度检验

图像可以直观地显示拟合程度的优劣，但是有时候“眼见”未必“为实”，通过严密的数学论证所得到的结论才更具有说服力，统计假设检验便是长期以来为人们所信赖的一种“论证”。

在假设检验中，我们先要设定原假设和备选假设：

H_0：数据来源于某个给定的总体；

H_1：数据并非来源于给定的总体。

针对原假设的不同，将有两种处理的方式。如果原假设中给出了完整的模型（如：均值为0，方差为1的标准正态分布），检验临界值可以较为容易地得出；另一种情况是原假设仅仅指明了模型的类型，而模型中仍含有待定的参数。如果模型的参数是通过样本数据估计得出（如用第九章的参数估计方法），这时的检验统计量要比事先给定模型时的统计量要小。这是因为参数估计时，我们会尽量使得分布函数与实际数据更接近，这会导致假设检验变成了近似。通常统计量较大时容易拒绝原假设，因此这种近似增加了犯第二类错误的概率，同时减小了犯第一类错误的概率。

针对第二种情况，我们可以通过将样本随机分组的方式避免近似。将样本随机分为两部分，一部分进行参数估计，另一部分进行假设检验。当模型选定之后，又重新将所有数据用于参数估计。

在本节接下来的内容中，我们以构造不同的检验统计量为主线介绍四

种假设检验的方法：拟合优度检验、K－S 检验、Anderson－Darling 检验以及似然比检验。

10.3.1　χ^2 拟合优度检验

χ^2 拟合优度检验常用于离散分布的情况，如果是连续分布则需要把数据分成多个区间来考虑。χ^2 拟合优度检验实施的步骤如下：首先，选定任意 $k-1$ 个值使得 $t=c_0<c_1<c_2<c_3<c_4\cdots<c_k=\infty$，其中 t 为左截断点（如果没有截断则 $t=0$）。记 $\hat{p}_j=F^*(c_j)-F^*(c_{j-1})$ 为观测值落在 $(c_{j-1},\ c_j]$ 区间中的概率，要注意每组包括组上限，即左端是开区间、右端是闭区间。类似地，记 $p_{nj}=F_n(c_j)-F_n(c_{j-1})$ 为由经验分布得到的 $(c_{j-1},\ c_j]$ 区间中的概率。然后构造 χ^2 检验统计量为：

$$Q=\sum_{j=1}^{k}\frac{n(\hat{p}_j-p_{nj})^2}{\hat{p}_j} \tag{10.3.1}$$

其中 n 为样本量。若令 $E_j=n\hat{p}_j$ 为区间中观测值个数的期望值，并令 $O_j=np_{nj}$ 为区间中的实际观测个数。此时有

$$Q=\sum_{j=1}^{k}\frac{(E_j-O_j)^2}{E_j} \tag{10.3.2}$$

当观测值 n 的值充分大时，统计量 Q 的分布会收敛于自由度为 $k-1-m$ 的 χ^2 分布，m 为模型中待估参数的个数。如果计算得到的 Q 大于临界值 $\chi^2_\alpha(k-1-m)$，则拒绝原假设，表明原假设中的分布不能拟合样本数据。否则，无法拒绝原假设。这里 α 通常取 0.05。

在 χ^2 拟合优度检验中，一定要注意满足：样本容量 n 要足够大、E_j 不太小这两个条件。为提高模型估计的精度，通常认为 E_j 的值不小于 5，总体的样本数据不小于 50，否则需要将个数较少的组合并，以满足这个要求。经验表明，当 E_j 的值大致相等时，检验的效果是最优的。所以当对个体数据进行检验时，应该适当地分组，使检验的效果尽可能好。

【例 10－5】　设某保险人经营某车辆险，对过去所发生的 1 000 次理赔情况做了记录，平均理赔额是 2 200 元，又按赔付金额分作 5 档，各档中的记录次数如表 10－3 所示。试用 χ^2 拟合检验判断是否能用指数分布来拟合个体理赔额的分布。

表 10－3　　车险赔付频率统计

赔付额（元）	0～1 000	1 000～2 000	2 000～3 000	3 000～4 000	4 000～5 000	5 000 元以上
次数	200	300	250	150	100	5

解： 先设个别理赔额 X 服从指数分布，使用矩估计或极大似然估计可以估计出 $\hat{\lambda}=2\ 200$。

下面计算 E_i：

$$E_i = nP_\lambda(X \in (C_{i-1}, C_i]) = n\ (F\ (C_i;\ \lambda) - F\ (C_{i-1};\ \lambda)) \qquad (10.3.3)$$

例如，在 2 000 ~ 3 000 组内的 E_3 为：

$$1\,000 \times \int_{2\,000}^{3\,000} \frac{1}{\lambda} e^{-x/\lambda} dx = 1\,000 \times (e^{-2\,000/\lambda} - e^{-3\,000/\lambda}) = 1\,000 \times 0.1472 = 147.2$$

类似地，可以计算得到其他组的平均次数：

$$E_1 = 365.3,\ E_2 = 231.8,\ E_4 = 93.4,\ E_5 = 59.3,\ E_6 = 103$$

因此，χ^2 统计量的值为：

$$Q = \frac{(200 - 365.3)^2}{365.3} + \frac{(300 - 231.8)^2}{231.8} + \cdots + \frac{(5 - 103)^2}{103} = 322.13$$

χ^2 分布的自由度为 $k - 1 - m = 6 - 1 - 1 = 4$，查表得，$\chi^2_{0.05}(4) = 9.487$，因此，$Q \geqslant \chi^2_{0.05}(4) = 9.487$，故应拒绝原假设，即选择指数分布不恰当。■

【**例 10 - 6**】 使用表 10 - 1 的数据，用拟合优度检验的方法验证可否使用指数分布进行拟合。

解：在 10.2 节中，我们曾经用分布函数与经验分布函数图像比较的方法进行了初步直观的判断，认为可以用指数分布进行拟合，现在用拟合优度检验方法进一步判断。

首先将数据分为 6 档，然后计算相应的 χ^2 值，得到表 10 - 4。

表 10 - 4　　拟合优度检验 χ^2 值

区间	$\hat{p}$	期望值	观察值个数	χ^2
50 ~ 150	0.1172	2.227	3	0.2687
150 ~ 250	0.1035	1.966	3	0.5444
250 ~ 500	0.2087	3.964	4	0.0003
500 ~ 1 000	0.2647	5.029	4	0.2105
1 000 ~ 2 000	0.2180	4.143	3	0.3152
2 000 ~ ∞	0.0880	1.672	2	0.0644
合　计	1	19	19	1.4034

自由度为 $k - 1 - m = 6 - 1 - 1 = 4$，$\alpha$ 水平为 5% 时，χ^2 临界值为 9.4877，P 值为 0.8436。因此可以得到结论：可以用指数分布来拟合此数据。■

10.3.2　K - S 检验

K - S 检验用来检验单一样本是否来自某一特定分布，比如检验一组数据是否为正态分布。这个检验的思想是：虽然 $Y_1, Y_2, \cdots, Y_n$ 的分布未知，但根据大样本理论，$Y_1, Y_2, \cdots, Y_n$ 的经验分布函数 $F_n(x)$ 在某种意义下收敛于其真实的分布，所以可以把 $F_n(x)$ 与所假设的分布函数 $F^*(x)$ 作比较，看它们

是否吻合。如果它们不能很好地吻合，就拒绝 H_0，即未知的真实分布函数不是由 $F^*(x)$ 给定的。由于经验分布 $F_n(x)$ 和分布 $F^*(x)$ 都是 x 的函数，所以要比较两者的差异需要一个合适的度量。Kolmogorov - Smirnov 检验提出一个最简单的度量，就是用 $F^*(x)$ 和 $F_n(x)$ 在垂直方向上的最大距离作为统计量。令 t 为左截断点（如果没有截断则 $t=0$），u 为右删失点（如果没有删失则 $u=\infty$）。这时检验统计量为：

$$D_n = \max_{t \le y \le u} | F_n(y) - F^*(y) | \tag{10.3.4}$$

需要注意的是，为确保 $F_n(x)$ 有定义，这个统计量只适用于个体数据，且要求 $F^*(x)$ 在对应区间上是连续的。由于经验分布函数 $F_n(x)$ 在每个数据点都有跳跃，因此需要将分布函数与跳跃前的值都进行比较。具体来说，如果已知一个样本观测值 x_1，…，x_n，则 D_n 为 $F^*(x)$ 与 $F_n(x)$ 差距的最大值为：

$$D_n = \max_{i=1,\cdots,n} \{ | F_n(x_{i-1}) - F^*(x_i,\hat{\theta}) | , | F_n(x_i) - F^*(x_i,\hat{\theta}) | \} \tag{10.3.5}$$

用 $Y=\sqrt{n}D_n$ 来表示被检验的实际偏差度量，其中 n 为样本数。对于小的 n 值，表 10－5 给出了 Y 在不同显著性水平下的临界值。

如果 $u<\infty$，临界值还应当取得更小。

图 10－7 K－S 统计量意义的图像展示

当 $n \to \infty$ 时，若 $F^*(x)$ 的函数形式完全给定，Y 的近似分布为：

$$P(Y>x) = 2\sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2x^2} \tag{10.3.6}$$

表 10－5 K－S 不同显著性水平下的 Y 的临界值

显著性水平（α）	临界值
0.2	1.07
0.1	1.22
0.05	1.36
0.01	1.63

若 $F^*(x)$ 的形式已知，参数由数据估计，偏差度量 $Y=\sqrt{n}D_n$ 将取得与前面结果不同的概率值，在这种情况下则需要对 Y 进行修正。例如，假定一个指数模型，θ 由数据估计而得，则

$$Y^* = \left(D_n - \frac{0.2}{n}\right)\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right) \tag{10.3.7}$$

为一个比未修正 $Y=\sqrt{n}D_n$ 更好的估计。关于这些修正的具体方法和临界值表，可见 Stephens, M. （1986）的论文。

【例 10－7】 我们仍然以例 10－2 中的车险数据（在 50 处截断）为例，我们依然用指数分布来拟合数据，并用数据来估计分布参数。最后使用K－S 检验来说明指数分布是否为合适的拟合分布。

解： 对于此数据，截断后分布 $F(x,\hat{\theta})$ 为：

$$F^*(x)=\frac{1-e^{-x/802.32}-(1-e^{-50/802.32})}{1-(1-e^{-50/802.32})}=1-e^{-(x-50)/802.32}$$

将经验分布函数和估计分布列表，见表 10－6。

表 10－6　　K－S 检验统计量计算结果

x	$F^*(x)$	$F_n(x-)$	$F_n(x)$	D
82	0.0391	0.000	0.0526	0.0391
115	0.0778	0.0526	0.1053	0.0275
126	0.0904	0.1053	0.1579	0.0675
155	0.1227	0.1579	0.2105	0.0878
161	0.1292	0.2105	0.2632	0.134*
243	0.2138	0.2632	0.3158	0.102
294	0.2622	0.3158	0.3684	0.1062
340	0.3033	0.3684	0.4211	0.1178
384	0.3405	0.4211	0.4737	0.1332
457	0.3979	0.4737	0.5263	0.1284
680	0.5440	0.5263	0.5789	0.0349
855	0.6433	0.5789	0.6316	0.0644
877	0.6433	0.6316	0.6842	0.0409
974	0.6839	0.6842	0.7368	0.0529
1193	0.7594	0.7368	0.7895	0.0301
1340	0.7997	0.7895	0.8421	0.0424
1884	0.8983	0.8421	0.8947	0.0562
2558	0.9561	0.8947	0.9474	0.0614
3476	0.9860	0.9474	1.000	0.0386

经过对最大值的简单观察便可得出：D 统计量为 0.134。$Y=0.134\sqrt{19}=0.5841$ 小于 5% 水平的临界值 1.36 说明指数分布是合适的分布。■

【例 10－8】 给出索赔额样本数据如下：29，64，90，135，182。原假设是样本来自指数分布。指数分布参数用矩估计的方法给出，计算 K－S 统计量的值。

解： 首先均值 $\hat{\theta}=\bar{x}=100$，则原假设的分布函数为：$F^*(x)=1-e^{-x/100}$。则 K－S 统计量与计算结果如表 10－7 所示。

表 10-7 例 10-8 的 K-S 统计量计算结果

x	$F_n(x)$	$F_n(x-)$	$F^*(x)$	$\lvert F_n(x)-F^*(x)\rvert$	$\lvert F_n(x-)-F^*(x)\rvert$
29	0.2	0	0.252	0.052	0.252
64	0.4	0.2	0.473	0.073	0.273
90	0.6	0.4	0.593	0.007	0.193
135	0.8	0.6	0.741	0.059	0.141
182	1.0	0.8	0.838	0.162	0.038

对表格最后两行的简单观察可以得出 K-S 统计量的值为 0.273。 ■

拟合优度检验与 K-S 检验都是采用实际频数与期望频数之差进行检验。它们之间的不同在于前者主要用于分组数据，而后者用于有计量单位的连续和定量数据。拟合优度检验虽然也可以用于定量数据，但是必须先将数据分组才能获得实际的观测频数；K-S 检验可以直接对原始数据的 n 个观测值进行检验，所以它对数据的利用较为完整。

10.3.3 Anderson-Darling 检验

K-S 检验只是简单地建立在经验分布 $F_n(x)$ 和模型分布 $F^*(x)$ 之间的最大偏差基础之上，T. W. Anderson 和 D. A. Darling 提出 de Anderson-Darling 检验给出了分布 $F_n(x)$ 和分布 $F^*(x)$ 之间偏差的更好度量。Anderson-Darling 统计量是 $F_n(x)$ 和分布 $F^*(x)$ 之间偏差的平方的加权期望值，权重是 $F_n(x)$ 方差的倒数，即

$$A^2 = n\int_t^u \frac{[F_n(x) - F^*(x)]^2}{F^*(x)[1 - F^*(x)]} f^*(x)dx \tag{10.3.8}$$

其中 t 和 u 的定义与 K-S 检验相同。注意当 x 接近于 t 或 u 时，分母很小，从而权重较大，因此这个统计量更加看重尾部的估计。这个积分很难计算，但是对于个体数据来说，记无重复的未被删失的数据点为 $t = y_0 < y_1 < \cdots < y_{k+1} = u$，则积分变为如下的求和形式：

$$\begin{aligned} A^2 = & -nF^*(u) + n\sum_{j=0}^{k-1}[1 - F_n(y_j)]^2\{\ln(1 - F^*(y_j)) - \ln(1 - F^*(y_{j+1}))\} \\ & + n\sum_{j=1}^{k} F_n(y_j)^2[\ln F^*(y_{j+1}) - \ln F^*(y_j)] \end{aligned} \tag{10.3.9}$$

若 $u = \infty$，则第一个求和项的最后一项等于 0。

当 $t = 0$，$u = \infty$ 时，可以证明式（10.3.8）与下面公式等价①

① 证明参见李晓林、孙佳美编著：《生命表基础》附录 A，中国财政经济出版社 2006 年版。

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\{\ln[F^*(x_i)\cdot S^*(x_{n-i+1})]\} \quad (10.3.10)$$

与式（10.3.8）相比，式（10.3.10）在计算 A^2 时更方便一些。显著性水平为10%、5%和1%的临界值分别是1.933、2.492和3.875。如果 $u<\infty$，这些临界值还应当更小一些。

【例10－9】 为便于方法间的比较，我们仍使用例10－2中的车险数据作为拟合的样本数据。

解： 根据式（10.3.9）对以上样本数据做列表运算，见表10－8。

表10－8　　Anderson－Darling检验统计量计算结果

j	Y_j	$F^*(x)$	$F_n(x-)$	求和项
0	50	0.000	0.000	0.0399
1	82	0.0391	0.0526	0.0388
2	115	0.0778	0.1053	0.0126
3	126	0.0904	0.1579	0.0332
4	155	0.1227	0.2105	0.007
5	161	0.1292	0.2632	0.0904
6	243	0.2138	0.3158	0.0501
7	294	0.2622	0.3684	0.0426
8	340	0.3033	0.4211	0.0389
9	384	0.3405	0.4737	0.0601
10	457	0.3979	0.5263	0.1490
11	680	0.5440	0.5789	0.0897
12	855	0.6433	0.6316	0.0099
13	877	0.6433	0.6842	0.0407
14	974	0.6839	0.7368	0.0758
15	1 193	0.7594	0.7895	0.0403
16	1 340	0.7997	0.8421	0.0994
17	1 884	0.8983	0.8947	0.0592
18	2 558	0.9561	0.9474	0.0308
19	3 476	0.9860	1.0	0.0141
20	∞	1.0	1.0	0
合计				1.0226

例如，表中求和项的第三个数据的计算公式是：

$$0.0126 = (1-0.10532)^2[\ln(1-0.0778)-\ln(1-0.0904) + 0.10532^2[\ln(0.0904)-\ln(0.0778)]$$

A^2 统计量等于 $-19(1)+19(1.0226)=0.4292$，小于5%的临界值

2.492，因此指数分布可以作为该车险数据的拟合分布。■

【例10－10】　对于例10－1中的样本数据，假设 $S^*(t)=\exp(-0.05t)$，利用式（10.3.10）计算 A^2 统计量的值并给出检验的结果。

解： 表10－9给出了统计量 A^2 的计算过程。

表10－9　　Anderson－Darling 检验统计量

j	t_j	$F^*(t_j)$	$S^*(t_{11-j})$	$(2i-1)\ \ln[F^*(t_j)\cdot S^*(t_{11-j})]$
1	3	0.13929	0.54881	－2.57120
2	4	0.18126	0.60653	－6.62347
3	5	0.22120	0.60653	－10.04345
4	6	0.29531	0.60653	－12.03812
5	7	0.29531	0.67032	－14.57757
6	8	0.32967	0.70469	－16.05626
7	9	0.39346	0.70469	－16.67605
8	10	0.39346	0.77880	－17.74165
9	11	0.39346	0.81873	－19.25721
10	12	0.45118	0.86070	－17.97207
合计				－133.55705

由式（10.3.10），$A^2=-10-\frac{1}{10}\sum_{j=0}^{10}(2i-1)\ln[F^*(t_j)\cdot S^*(t_{11-j})]$，则检验统计量的值可以由以上表格的求和式得出，为：$A^2=-10-\frac{1}{10}(-133.55705)=3.35571$。

A^2 统计量0.01的显著性水平下对应的值为3.857，则 $3.35571<3.857$，因此可以判断可以用该指数分布进行拟合。■

在使用这些拟合优度检验法时，需要注意样本容量对检验结果的影响。假设样本容量加倍但样本点的取值并没有太大变化（想象每个数据点都重复出现两次），在Kolmogorov-Smirnov检验中检验统计量不变，但临界值会变小；在Anderson-Darling和 χ^2 拟合优度检验中，检验统计量会加倍但临界值不会改变。这使得对于较大容量的样本，原假设更容易被拒绝。这也很容易理解，因为原假设本身就是错误的（由几个参数决定的某个分布函数就能够合理解释观测值表现的各种复杂现象，完全解释的可能性极小），但只有在样本容量足够大的时候，才会有可以令人信服的证据来说明这一点。

10.3.4　似然比检验

与前面三种检验不同的是，似然比检验考虑的是两个分布的比较。该检验的原假设和备选假设分别为：

H_0：数据来自服从 A 分布的总体；

H_1：数据来自服从 B 分布的总体。

为了能够进行正规的假设检验，A 分布必须是 B 分布的一种特殊情形，如指数分布相对于伽玛分布。具体来说，假定总体分布的密度函数（连续型）或概率分布函数（离散型）为 $f(x;\theta)$，其中 θ 为参数，且 $\theta \in \Theta$。似然比检验考虑如下假设检验问题：

H_0：$\theta \in \Theta_0$，

H_1：$\theta \in \Theta_1 = \Theta - \Theta_0$

其中 Θ_0 为 Θ 的一个子集。

对于给定的样本 $x_1, x_2, \cdots, x_n$，似然函数为 $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$。定义统计量

$$LR = \frac{\sup\{L(\theta):\theta \in \Theta\}}{\sup\{L(\theta):\theta \in \Theta_0\}} \tag{10.3.11}$$

称 LR 为似然比（*Likelihood Ratio*）。LR 的分子是参数没有被约束的似然函数最大值，分母是参数被约束时的最大值。显然有：$LR \geqslant 1$。

根据极大似然估计的性质，θ 的极大似然估计 $\hat{\theta}$ 是 θ 的相合估计，即：$\hat{\theta} \xrightarrow{p} \theta\ (n \to \infty)$。所以当假设 H_0 为真时，$\hat{\theta}$ 应以大的概率属于 Θ_0。这时，如果似然函数 $L(\theta)$ 具有连续性（大多数场合都是这样的），似然比 LR 应接近于 1。反之，LR 应该远大于 1。Wilks 在 1938 年证明了：在一定正则的条件下，$Y_n = 2\ln LR$ 在原假设下以 χ^2 分布为极限分布，参数为 $k-r$，k 为没有被约束的参数个数，r 为被约束的参数个数。若 $\Theta_0 = \theta_0$，θ_0 为已知参数，则 $r=0$。若 Y_n 大于置信水平为 α 临界值 c，则拒绝原假设，即认为 $\theta \in \Theta_1$。

【例 10－11】 一个来自指数分布总体 X 的样本包含 8 个数据：1.0，1.2，1.3，1.7，1.9，2.0，2.4，2.5。在 5% 显著性水平下进行如下似然比检验：

H_0：$\theta = 2$，H_1：$\theta \neq 2$

解：似然函数是 $L(\theta) = \prod_{i=1}^{8} f(x_i) = \prod_{i=1}^{8} \frac{1}{\theta} e^{-x_i/\theta} = \frac{1}{\theta^8} e^{-\sum x_i/\theta}$

在原假设下，$\theta_0 = 2$，则

$$L_0 = L(\theta_0) = L(2) = \frac{1}{2^8} e^{-\sum x_i/2} = 0.000003562$$

根据样本数据，对 θ 的极大似然估计是 $\theta_1 = \bar{x} = 1.75$，此时有：

$$L_1 = L(\theta_1) = L(1.75) = \frac{1}{1.75^8} e^{-\sum x_i/1.75} = 0.000003814$$

似然比统计量是 $T = 2\ln \frac{L_1}{L_0} = 0.14$，且自由度是 1，此时对应 5% 显著性

水平下的临界值是 $\chi^2_{0.05}$ (1) =3.84 > T，因而此时无法拒绝原假设。 ■

【例 10-12】 假设机动车辆险中私人小汽车的赔付额服从指数分布。表 10-10 的数据是两个不同品牌的汽车发生的赔付额。根据这批数据能否断定这两个品牌的汽车的赔付额服从同一指数分布？（取 $\alpha=0.05$）？

表 10-10 汽车品牌与赔付额

品牌	赔付额（百元）										
A	74	57	48	29	502	12	70	21	486	59	27
B	55	320	56	104	220	239	47	246	176	182	33

解： 假定这两个品牌的汽车的赔付额分别服从参数为 θ_1 和 θ_2 的指数分布。根据题意，原假设为：H_0：$\theta_1=\theta_2$。对参数作一变换，记：$\theta_1=\theta$，$\theta_2=\theta+d$，则上述假设就转换为：

$$H_0: d=0$$

用 x_1，x_2，$x_3,\cdots,x_{n1}$和 y_1，y_2，$y_3,\cdots,y_{n2}$分别记两组数据，其中 $n_1=11$，$n_2=11$。样本均值分别记为：$\bar{x}$，$\bar{y}$，记样本总均值为 $\bar{z}$，则 $\bar{x}=109.5$，$\bar{y}=152.5$，$\bar{z}=130.1$，无约束参数的极大似然估计值为：$\hat{\theta}_1=109.5$，$\hat{\theta}_2=152.5$。

当假设 H_0 成立时，参数的极大似然估计值为：

$$\tilde{\theta}_1=\tilde{\theta}_2=\bar{z}=130.1$$

对数似然函数为：

$$l(\theta_1,\theta_2)=-\left(\sum_i x_i/\theta_1+\sum_j y_j/\theta_2\right)-n_1\ln\theta_1-n_2\ln\theta_2 \quad (10.3.14)$$

似然比统计量为：$2\ln LR=2[l(\hat{\theta}_1,\hat{\theta}_2)-l(\tilde{\theta}_1,\tilde{\theta}_2)]=0.6316$。

由于 $2\ln LR$ 渐进分布为 $\chi^2_{0.95}$ (1)，在显著性水平 $\alpha=0.05$ 下，$\chi^2_{0.95}$ (1) =3.841，因此 $2\ln LR<\chi^2_{0.95}$ (1)，故接受原假设，即认为两个品牌的汽车的赔付额服从同一指数分布。 ■

在使用似然比检验时，原假设的分布应是备选假设分布的特殊情形。当原假设为备选假设的极限情形时（不是一种特殊情形），似然比检验仍然适用，此时检验统计量服从混合 χ^2 分布。但是当备选假设分布仅比原假设分布具有更多的参数时，似然比检验是不合适的。例如，两参数的对数正态模型完全有可能比三参数的 Burr 模型的对数似然函数值大，这样就会产生负的检验统计量，进而无法使用 χ^2 分布。

§10.4 最优模型的选择

在对众多备选模型进行了分布图像的比较和相关的拟合检验之后，剩

下的任务是作出最终的选择。在筛选时，需要注意节俭原则和相对最优原则。

1. 节俭原则：对于具备同样条件的模型，在没有充分的理由支持复杂模型时，简单模型是更好的选择。这是因为虽然复杂模型对已知数据拟合得很好，但是并不能保证对观测总体的拟合同样好。这与大多数假设检验的方法是一致的：如果没有十分有力的证据，就不要拒绝原假设。

2. 相对最优原则：无论选择哪个模型，它都是对实际情况的一种近似。模型选择目标是找到一个能够解决问题的足够好的模型，而足够好的定义将取决于具体的应用。如果尝试了足够多的模型，即使所有的尝试过的模型拟合效果都不足够理想，但总会有一个模型是相对不错的，虽然这并不能提高我们对总体性质的认识。

因此，我们在选择模型时要注意在简洁性与精确性之间权衡。只要可能，应尽量选择简单的模型，且还要根据实际情况来限制备选模型的范围。例如，选择便于计算的模型来厘定费率，或者选择厚尾分布来估计再保险损失等。这需要建模者有丰富的经验才能缩减备选模型的范围。

本节内容将分别介绍两类模型选择标准，第一类标准是基于建模者的主观判断，而第二类标准更为规范，使得大多数情况下所有人都会通过分析得出相同的结论，因为后者的结论不是由图形和表格的直观印象得来，而是通过对数据的定量分析得到的。

10.4.1 主观判断法

主观判断法是指根据精算师的经验和实际背景来确定备选模型，筛选模型。主要从以下三个方面来考虑：

第一，可以根据本章介绍的各种图表工具（或者基于图形的表格）进行基本的判定。例如利用经验分布图（卵形图）、直方图、核密度图等，比较经验分布的对称性、偏斜程度以及分布的扁平程度是否与备选的分布函数一致。有时还要比较经验分布的尾部特征，是否有很长的尾部，是否存在厚尾。在进行真正的分析时可以集中在关系重要的方面，例如，有时候对尾部概率的拟合非常重要，而有时候众数的匹配更重要。即便是评分法，也可以运用一幅具有说服力的图形来支持选定的模型。

第二，根据行业背景和使用目的来选择模型。有时候评分法认为某个模型的拟合效果可能比其他模型要好很多，但是从计算方便的角度来说，可能另一个模型虽然拟合效果稍微差些，但能使某些计算大为简化。比如，1941 年的美国保险委员会标准寿命表（Commissioners' Standard Ordinary Mortality Table，CSO）在许多年龄段都采用了 Makeham 分布。在计算能力有限的年代，这种分布使得对联合生存相关值的计算大大简化。类似地，

在一般责任保险和机动车第三责任险中，精算师普遍使用截断帕累托作为损失的分布，因为该分布具有某些良好的性质。那么当采用其他分布时，可能就需要提出比通常情况更充足的证据才行。

第三，分布可能完全由具体情形决定。例如，某健康险合约为女性投保人提供每年至多两次乳腺癌和宫颈癌的两癌筛查体检，而投保个体每年会有两次独立的选择，决定是否进行体检，如果每次决定体检的概率是 q，则体检次数必然服从 $m=2$ 的二项分布。

10.4.2　评分法

当存在多种分布可以用来拟合样本数据的分布时，我们应该选择一个“最优”的分布作为样本数据的分布。一种简单的方法是为每个模型打分，得到最优分数的模型将被选择。常用的打分方法及判断标准如表 10－11 所示。

表 10－11　　评分方法及判断标准

方　　法	分数依据	判断标准
似然函数法	极大似然函数值	越大越好
χ^2 拟合优度检验法	χ^2 统计量的值	越小越好
χ^2 拟合优度 p 值法	p 值：$p=P\ (\chi^2>Q)$	越大越好
K－S 检验法	K－S 检验统计量	越小越好
Anderson－Darling 检验法	Anderson－Darling 统计量	越小越好

在使用这几种判断标准时，需要考虑参数的个数对分值的影响。拟合程度也可能随着模型复杂度的增加而增加。以似然函数法为例，当比较指数模型和伽玛模型对某样本的拟合程度时，因为指数分布是伽玛分布的特殊情形，因此伽玛模型的似然函数值不小于指数模型的似然函数值，得到伽玛分布总会胜过指数分布的结论。这违背了节俭原则。对于表 10－11 中的五种方法，除 χ^2 拟合优度 p 值法不受模型复杂度的影响，其他方法都存在节俭原则上的不足。其原因是，随着模型复杂程度的增加，检验的自由度减小，所以复杂的模型也可能有较小的 p 值。因此从节俭原则的角度出发，χ^2 拟合优度 p 值是最优的评分方法。

如果用似然函数作为评分的标准，为了不违背节俭原则，有两种解决的方法。一种是进行似然比检验，它适合于一个模型是另一个模型的特例时（例如伽玛模型和变换伽玛模型），或者一个模型是另一个模型的极限情形时（例如 Bull 分布和韦伯分布）。可以首先从单参数的模型（有极大似然函数值的模型）开始，然后和两参数模型比较，如果两参数模型的极大似然函数值增加了至少 1.92（将这个增量加倍就得到临界值的增量

3.84)，则采用两参数模型。再考虑三参数模型，如果和两参数模型比较，仍然需要 1.92 的增量，而如果第一步保留了单参数模型，则三参数模型必须有 3.00 的增量才能被采用（因为该检验的自由度是 2）。若要增加三个参数，则需要 3.91 的增量，四个参数需要 4.74 的增量，等等。似然比检验也可以在模型之间没有包含关系的情形中使用，但是可能无法认为这种情形下仍然在进行似然比检验。除特殊情形外，似然比检验也面临着其他假设检验相同的问题。当样本容量加倍时，对数似然函数值也会加倍，使得参数个数更多的模型更容易被选择，这将使得违背节俭原则的结果出现。

另一种是在引入新的参数时有一定的惩罚，可考虑 Schwarz Bayesian 准则（SBC）或 AIC 信息准则之类的判定准则。SBC 准则衡量模型时将对数似然函数值减$(r/2)\ln n$，即 $\mathrm{SBC}=\ln L-(r/2)\ln n$，其中 r 是待估参数的个数，n 是样本容量。这样一来，只有对数似然函数增加 $0.5\ln n$ 才能增加一个参数，样本容量越大，要求的似然函数增量就越大，但这个要求的增量并非随样本容量比例增长。AIC 信息量准则（Akaike information criterion）也是权衡所估计模型的复杂度和此模型拟合数据的优良性的一种标准。在一般的情况下，AIC 可以表示为 $\mathrm{AIC}=2r-2\ln L$。在 AIC 准则下，通常选择模型应是 AIC 值最小的那一个。

一定要注意，通过严格的计算选择模型的评分法，得到的结果并不一定一致。这时，必然要通过主观判断来决定最终的选择。

【例 10－13】 请比较韦伯分布和指数分布哪个更适合于描述例 10－2 中的数据。

解： 对例 10－2 中的数据分别计算指数分布和韦伯分布的列表中统计量的值，相应的计算结果见表 10－12（在 50 处截断）。

表 10－12　　例 10－13 数据不同评分方法的计算结果

标　准	指数分布	韦伯分布
K－S 值	0.1340	0.0887
A－D 值	0.4292	0.1631
χ^2 值	1.4034	0.3615
P 值	0.8436	0.9481
极大似然值	－146.063	－145.683

对于以上的结果，可以发现韦伯分布都有一定的优势。因此在备选模型为指数分布和韦伯分布时，我们理应选择韦伯分布。■

【例 10－14】 我国某保险公司 1996 年的 35 072 辆投保车辆索赔次数统计结果如表 10－13 第 1、2 列所示。试分析索赔次数的分布。

解：首先可以按照近似公式（4.3.20）计算 $k\frac{n_k}{n_{k-1}}$ 的值（如表 4-3 第 3 列所示），然后绘出其图形，见图 10-8。

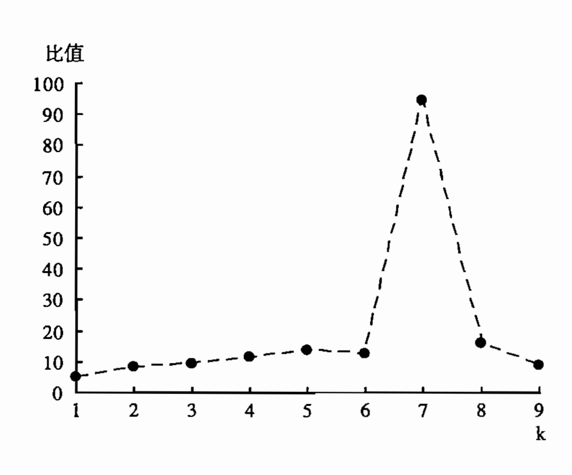


图 10-8　函数 $k\frac{n_k}{n_{k-1}}$ 的曲线

表 10-13　车辆索赔次数数据

索赔次数 k	保单数 n_k	$k\frac{n_k}{n_{k-1}}$
0	27 141	
1	5 789	4.69
2	1 443	8.02
3	457	9.47
4	155	11.79
5	56	13.84
6	27	12.44
7	2	94.5
8	1	16
9	1	9
≥10	0	
总计	35 072	

从图 10-8 中可以看出，除了 $k=7$ 这个点之外，其余各点近似呈一条直线，并且斜率为正，因此可以考虑用负二项分布来拟合索赔次数。当然，斜率为正也必须通过假设检验才能确定。我们初步判断可以采用负二项分布拟合，但是不能排除泊松分布。因为直线斜率较低。

首先估计泊松分布。由表 10-13 的数据可以算出样本均值和方差分别是 $\bar{x}=0.3176$，$s^2=0.4913$，以样本均值作为泊松分布的参数，可以得出实际观测数和拟合频数的比较，如表 10-14 所示。

表 10-14　实际观测值和用 P（0.3176）拟合结果比较

	车辆数		
索赔次数	观测值	拟合值	误差
0	27 141	25 528.69	1 612.31
1	5 789	8 107.91	-2 318.91
2	1 443	1 287.54	155.46
3	457	136.31	320.69
4	155	10.82	144.18
5	56	0.69	55.31
6	27	0.04	26.96
7	2	0	2

续表

	车辆数		
索赔次数	观测值	拟合值	误差
8	1	0	1
9	1	0	1
≥10	0	0	0
总计	35 072	35 072	0

其次，采用二元结构模型，用两个泊松分布的混合再次拟合，结果见表 10－15。

表 10－15　实际观测值、利用一元结构函数以及二元结构函数拟合结果的比较

索赔次数	车辆数				
	观察值	拟合值		误差	
		$N \sim P$（0.3176）	$N \sim 0.8876P$（0.1719）$+0.1124P$（1.4694）	一元结构函数	二元结构函数
0	27 141	25 528.69	27 120.34	1 612.31	20.66
1	5 789	8 107.91	5 838.70	－2 318.91	－49.7
2	1 443	1 287.54	1 366.37	155.46	76.63
3	457	136.31	501.74	320.69	－44.74
4	155	10.82	177.12	144.18	－22.12
5	56	0.69	51.81	55.31	4.19
6	27	0.04	12.68	26.96	14.32
7	2	0	2.66	2	－0.66
8	1	0	0.49	1	0.51
9	1	0	0.08	1	0.92
≥10	0	0	0.01	0	－0.01
总计	35 072	35 072	35 072	0	0

从表 10－15 可以看出，用混合泊松分布拟合的误差明显降低。

最后，结构函数采用伽玛函数，参数分别为 α 和 θ，有

$$\begin{cases} E(N)=\alpha\theta=\bar{x} \\ Var(N)=\alpha\theta+\alpha\theta^2=s^2 \end{cases}$$

因此 α 和 θ 的矩估计值分别为：

$$\begin{cases} \hat{\alpha}=\bar{x}^2/(s^2-\bar{x})=0.5807 \\ \hat{\theta}=(s^2-\bar{x})/\bar{x}=0.5469 \end{cases}$$

即负二项分布的参数估计值分别为：

$$\hat{r}=\hat{\alpha}=0.5807$$

$$\hat{p}=\frac{1}{1+\hat{\theta}}=0.6464$$

再次拟合的结果见表 10－16。

表 10－16　实际观测值、利用一元结构函数、二元结构函数以及伽玛结构函数拟合结果的比较

索赔次数	车辆数						
	观察值	拟合值			误差		
		$N\sim P$ (0.3176)	$N\sim 0.8876P$ (0.1719) $+0.1124P$ (1.4694)	$N\sim NB$ (0.5807, 0.6464)	一元结构函数	二元结构函数	伽玛结构函数
0	27 141	25 528.69	27 120.34	27 223.0	1 612.31	20.66	－82
1	5 789	8 107.91	5 838.70	5 589.2	－2 318.91	－49.7	199.8
2	1 443	1 287.54	1 366.37	1 561.8	155.46	76.63	－118.8
3	457	136.31	501.74	475.0	320.69	－44.74	－18
4	155	10.82	177.12	150.3	144.18	－22.12	4.7
5	56	0.69	51.81	48.7	55.31	4.19	7.3
6	27	0.04	12.68	16.0	26.96	14.32	11
7	2	0	2.66	5.3	2	－0.66	－3.3
8	1	0	0.49	1.8	1	0.51	－0.8
9	1	0	0.08	0.6	1	0.92	0.4
≥10	0	0	0.01	0.3	0	－0.01	－0.3
总计	35 072	35 072	35 072	35 072	0	0	0

从表 10－16 可以看出，二元混合泊松分布对前三个点的拟合效果较好，而负二项分布则对尾部的拟合效果较好。

【例 10－15】 表 10－17 的事故数据，总共有 67 856 个机动车事故保单。试选择一个合适的分布描述该数据。

表 10－17　机动车事故保单数据

索赔次数	0	1	2	3	4
频数	63 232	4 333	271	18	2

解： 由表 10－17 可以看出，索赔次数的分布比较集中，只包括 0，1，2，3，4 五种情况，并且索赔次数为 0 次的保单数最多，占总保单数的 93.19%；而且索赔次数越高，保单数越少。再具体研究索赔次数数据的特征，几个相关的统计量如表 10－18 所示。

表 10－18　　索赔次数的相关统计量

最小值	最大值	均值	方差
0	4	0.07275701	0.07739737

从上面数字特征可以看出，索赔次数数据的方差要大于均值，在常用的（a，b，0）分布类中，负二项分布具有这种性质，而且负二项分布具有两个参数，因此比泊松分布在拟合上会更加灵活。

通过索赔次数的频数表可以看出，索赔次数为 0 次的保单数最多，占总保单数的 93.19%。考虑到零膨胀现象，因此下面对零点的概率做适当的调整，使它正好等于一个特定值，同时调整其他非零点上的概率，使所有概率值和等于 1。

表 10－19 对泊松、负二项、零修正的泊松、零修正的负二项作了综合的比较。从表 10－19 中可以看出，从拟合优度检验的 p 值来看，负二项分布的效果是最好的。从似然值上来看，零修正的负二项分布的似然值最大，效果有一定提高。但是从 SBC 准则看，负二项分布是最优选择。

表 10－19　　四种不同分布的比较

索赔数	观测值	用于拟合的分布			
		泊松	负二项	零修正的泊松	零修正的负二项
0	63 232	63 094.323	62 333.100	63 230.000	63 230.000
1	4 333	4 590.544	4 321.293	4 321.707	4 330.303
2	271	166.998	276.273	286.220	270.228
3	18	4.050	17.209	12.637	18.945
4＋	2	0.075	1.125	5.436	6.524
参数		$\lambda=0.072757$	$r=1.1560670$ $p=0.9407908$	$\lambda=0.132457$	$r=0.4594147$ $p=0.9144807$
χ^2		177.153919	0.82193845	5.29	3.1881106
自由度		3	2	3	2
p 值		<0.01	0.6630	0.1520	0.3635
对数似然		－18 101.5	－18 049.68	－18 052.2	－18 049.47
SBC		－18 107.1	－18 060.8	－18 063.3	－18 066.2

习　题

1. 用 200 份赔付数据拟合一个帕累托分布，给定：

（1）对应的极大似然估计是 $\hat{\alpha}=1.4$ 和 $\hat{\theta}=7.6$；

（2）以极大似然估计值算得的对数似然函数值是 -817.92；

（3）$\sum \ln(x_i + 7.8) = 607.64$。

使用似然比检验对原假设 $\alpha = 1.5$ 和 $\theta = 7.8$ 进行检验。

2. 表 10-20 是一个包含 15 个损失数据的样本。已知损失额大小服从 $(0, \theta)$ 区间上的均匀分布。记 E_j 是第 j 个区间上的损失次数的期望值，O_j 是第 j 个区间上损失次数的实际观测值。试通过最小化 $\sum_{j=1}^{3} \frac{(E_j - O_j)^2}{E_j}$ 来估计 θ。

表 10-20

损失额大小所在区间	观察到的损失值
(0, 2]	5
(2, 5]	5
(5, ∞)	5

3. 假设赔付的大小服从指数分布。随机选取 5 个赔付样本 31、66、85、135、180。使用矩估计获得指数分布的参数。求对应的 Kolmogorov - Smirnov 检验统计量。

4. 将 365 天每日观察到的索赔数的数据汇总如表 10-21 所示。

表 10-21

每天索赔数	0	1	2	3	4 +
对应天数	50	122	101	92	0

首先结合以上数据用极大似然估计拟合一个泊松模型，然后将数据按照每天索赔数重新分为 4 组：0、1、2、3 +。对原假设“索赔数服从泊松分布”进行 χ^2 拟合优度检验，结果如何？

5. 给定以下包含 30 个数据的汽车索赔额数据：54、140、230、560、600、1 100、1 500、1 800、1 920、2 000、2 450、2 500、2 580、2 910、3 800、3 800、3 810、3 870、4 000、4 800、7 200、7 390、11 750、12 000、15 000、25 000、30 000、32 300、35 000、55 000。原假设为索赔额的分布服从一个分位数如表 10-22 所示的连续分布 $F(x)$。

表 10-22

x	310	500	2 498	4 876	7 498	12 930
$F(x)$	0.16	0.27	0.55	0.81	0.90	0.95

检验时在保证每组至少有 5 个数据的前提下分成尽可能多的组，计算 χ^2 拟合优度检验统计量。

6. 设用来拟合的模型是正确的，则当样本量逐渐增大至无穷时，Kolmogorov - Smirnov 检验统计量、Anderson - Darling 检验统计量和 χ^2 拟合优度检验统计量的大小将会如何变化？

7. 给定下列观测值的一个样本：0.1、0.2、0.5、0.7、1.3。为了检验对应的概率密度函数为

$$f(x)=\frac{4}{(1+x)^5},\ x>0$$

这个假设，计算相应的 Kolmogorov – Smirnov 检验统计量。

8. 设原假设为给定的数据来自一个已知分布 $F(x)$，如表 10 – 23 所示。求相应的 χ^2 拟合优度检验的结果。

表 10 – 23

区间	$F(x_i)$	实际观测数目
$x<2$	0.035	5
$2\leqslant x<5$	0.130	42
$5\leqslant x<7$	0.630	137
$7\leqslant x<8$	0.830	66
$8\leqslant x$	1.000	50
合计		300

9. 来自服从韦伯分布的总体的样本如下：595、700、789、799、1 109。已知在 θ 和 τ 的极大似然估计点，$\sum \ln(f(x_i))=-33.05$。当 $\tau=2$ 时，θ 的极大似然估计是 816.7。用似然比检验做一下检验 H_0：$\tau=2$，H_1：$\tau\neq 2$。求检验结果。

10. 观测到某险种的赔付额为：400、1 000、1 600、3 000、5 000、5 400、6 200。假设数据服从 $\theta=3\ 300$ 的指数分布，通过 $p-p$ 图和 $D(x)$ 图进行拟合优度检验。令 $(s-t)$ 表示赔付额 3 000 在 $p-p$ 图上的坐标，求 $(s-t)-D$（3 000）。

11. 来自总体 X 的包含 12 个数据的样本为：7、12、15、19、26、27、29、29、30、33、38、53。用于拟合的模型是参数为 $\hat{\theta}=30$ 的指数分布。基于上述数据画出 $p-p$ 图，找出在区间 [0，1] 上，该图像与 $y=x$ 的图像之间垂直方向上的最大离差（即求 max $\left|F^*\left(x_j-\frac{j}{13}\right)\right|$）。

12. 一个来自服从参数 $\hat{\theta}=15$ 的指数分布的总体的样本包含 8 个数据：3、4、8、10、12、18、22、35。根据以上数据，求 10% 显著性水平下 Kolmogorov – Smirnov 检验的结果和 10% 显著性水平下 Anderson – Darling 检验的结果。

13. 一组保单数为 50 的风险集的索赔数以如表 10 – 24 以下分组数据形式给出。记“H_0：各风险的索赔数服从 0、1、2、3、4 上的离散均匀分布”。

表 10 – 24

索赔数	0	1	2	3	4
保单数	7	10	12	17	4

（1）使用χ^2拟合优度检验去检验这个原假设；

（2）将2个相邻的组合并，则哪种合并方式会得到与上一题不同的结论？

14. 使用参数为$\hat{m}=7$和$\hat{q}=0.289$的二项分布拟合上题数据，使用χ^2拟合优度检验去检验拟合的优劣。并与第13题的结果进行基于p值的优劣对比。

15. 一个随机抽取的样本包括100个数据，用指数分布拟合时，以极大似然估计去求分布的参数，此时极大化的似然函数值为-159.4。继续用伽玛分布拟合这组数据，如果根据似然比检验，伽玛分布的拟合效果在5%显著性水平下优于指数分布的话，则用极大似然估计求伽玛分布模型的参数时，最大化的似然函数值应在什么范围内？

16. 一个来自总体X的样本包含12个数据：7、12、15、19、26、27、29、29、30、33、38、53。

（1）假设数据在32处删失，并使用参数为$\hat{\theta}=25$的指数分布拟合这组数据。求对应的Kolmogorov－Smirnov检验统计量。

（2）使用参数为$\hat{\theta}=30$的指数分布拟合这组数据。求对应的Anderson-Darling检验统计量。

17. 40个损失数据以千为单位被汇总如表10－25所示。

原假设为“损失额（以千为单位）的分布服从密度方程$f(x)=x^{-2}$，$x>1$”。对原假设进行χ^2拟合优度检验，求对应检验统计量的大小。

表10－25

区间（1 000）	损失数	区间总损失（1 000）
(1, 4/3]	16	20
(4/3, 2]	10	15
(2, 4]	10	35
(4, ∞)	4	20

18. 对模型A和B进行似然比检验。已知A是单参数模型，B不止一个参数。进行完参数估计后，两模型的对数似然函数值分别是$l_A=-280$和$l_B=-276$。如果检验的结果是在5%显著性水平下B优于A，则B的参数最多能有几个？

第三篇　模型的调整和随机模拟

第十一章　修匀理论

学习目标

- □ 了解修匀的基本原理，熟悉拟合检验和光滑检验的几种度量方法
- □ 了解表格数学修匀的常见方法，熟练使用 M－W－A 修匀法、Whittaker 修匀法和 Baysian 修匀法来修正初始估计值
- □ 熟悉参数估计的几种方法，熟悉分段参数修匀法和光滑连接修匀法
- □ 能够灵活运用最小二乘修匀法、极大似然函数法、样条修匀法来修正初始估计值

§11.1　修匀法概述

11.1.1　修匀的定义

本书的前面部分都是基于概率统计的，即根据观察数据，利用统计方法对未知参数进行估计，得到初始估计值等。但是由于样本信息不全或估计方法不科学等各种原因，估计结果往往存在一些明显的偏差。例如，一般认为死亡率应该是关于年龄连续递增的，但是在估计死亡率时，对每个年龄阶段的死亡率是单独估计的，很有可能会出现死亡率不满足关于年龄递增性，而且得到的是一个个离散值。如果有确切的理由对这种估计表示怀疑，是否可以根据所取得的估计值，获取更合理、更有效的估计？这些都是修匀学所需解决的问题。

Miller 把修匀叙述成：“修匀是这样一种可靠的方法，根据一个连续变量的不规则的观察序列，用这种方法，可得到一个光滑的、有规则的修正序列，与观察值序列相和谐。”①

为什么我们不把初始估计值看成是那些要去估计的未知值的最佳估计呢？如果我们采用的估计过程是合理的、无可非议的，为什么还要去改变

① 这句话来自 Dick London 著、徐诚浩译：《修匀数学》，上海科学技术出版社 1996 年版，第 1 页。

那些估计呢？答案在于数据本身。每个初始估计值都是某个特定序列的一个元素，在这些元素之间我们怀疑存在某种很强的关系。例如死亡率，我们认为它应该是关于年龄的连续变化、递增的光滑曲线。而在大多数情形下，这个序列的每个元素是彼此独立得到的，也就是说，不承认“相邻变化率之间有关系”。而当我们修匀这些初始估计值时，这些关系是被承认存在的，且被反映在它们的修正估计上。因此，所谓的**修匀就是这样的一种过程，它利用某种先验信息把初始估计值修正为新值，并把新值作为未知真实值的较好的估计。**

修匀在精算实务中一个重要的应用就是编制生命表。当用生存模型的理论得到初始生命表后，这些序列未必具有连续变化（光滑）、递增等体现序列内在联系的特点，利用先验的观点——光滑、递增，对原表进行修匀后，便得到最终生命表。

先验信息在本章中主要有两种，一种是通常的物理函数都具有某种内在联系，如函数连续、一阶可导或高阶可导（光滑）、递增等；另一种是先验的统计信息，如历史上已经得到的关于死亡率的分布。

为了严格地描述修匀，引入下列符号：

x——下标（常指年龄）；

t_x——待估参数（常指死亡率）；

U_x——t_x 的估计量，它是一个随机变量，当它作为死亡率的估计量时，可表示为二项比例$\frac{N_x}{n_x}$，而 N_x服从参数为 n_x 和 t_x 的二项分布 $B(n_x, t_x)$；

n_x——样本容量或暴露数；

u_x——U_x 的实现值，是 t_x 的初始估计值；

v_x——t_x 的修匀值；

V_x——以 v_x 为观察值的随机变量；

w_x——修匀表达式中的权序列；

E_x——估计误差随机变量；

e_x——E_x 的实现值。

它们的一般关系为：

$$u_x = t_x + e_x$$

$$U_x = t_x + E_x$$

设 G 是某个修匀过程（修匀算子），它作用在初始估计序列 u_x 上，我们可以认为 G 分别作用在分量 t_x 和 e_x 上，它保持 t_x 分量基本上不变化，而有效地减少误差分量 e_x，则

$$v_x = G(u_x) = G(t_x) + G(e_x) = t_x + e_x'$$

其中 $e'_x = G(u_x) - t_x$ 称为修匀误差。修匀的目的就是使 $|e'_x| < |e_x|$，从而

使 v_x 较 u_x 更接近于 t_x。

11.1.2 拟合检验和光滑性检验

在关于修匀的定义中，对于修匀估计值，有两个重要的要求：首先，修匀值不应离初始（观察）值太远，即这两者之间应有一定的拟合度；其次，根据先验假设，真实序列应该是连续变化的，所以修匀值应该具有一定的光滑度。因此在修匀过程中需要对修匀结果进行拟合检验和光滑性检验。

对于光滑性检验，传统的方法是计算这些修匀值的某些阶的有限差分。当那些差分的某个阶（最常用的是 3 阶和 4 阶）是“较小值”时，我们就认为得到了某个光滑度。例如采用 $S=\sum_i(\Delta^4 v_i)^2$ 作为光滑性的一个度量，这里 v_i 是指标 i 处的修匀值。这种光滑性度量，暗示着力求使修匀值分布在一条三次曲线上。当然，不同情况下的光滑度量是不同的。原则上，这种光滑度量应符合光滑性先验观点。

根据拟合的概念，可以给出一种衡量拟合程度的表达式：

$$F_1=\sum_x w_x(v_x-u_x) \tag{11.1.1}$$

这里的权数 w_x 所反映的是 x 岁的样本容量（暴露数）对总偏差的影响，如 $w_x=n_x$ 或 $w_x=n_x/\sum_y n_y$。F_1 越接近于 0，说明拟合程度越好。然而式（11.1.1）有个明显的缺陷是，当正或负的加权偏差同时出现但相互抵消时，一个很粗劣的拟合也可能使 $F_1=0$。鉴于这一点，可考虑下面衡量拟合程度的表达式：

$$F_2=\sum_x w_x(v_x-u_x)^2 \tag{11.1.2}$$

或者加权偏差的一阶矩的和

$$F_3=\sum_x xw_x(v_x-u_x) \tag{11.1.3}$$

尽管 F_1 和 F_3 有明显的缺陷，但是在处理死亡数据时，仍经常被使用，这是因为它们有如下的直观含义：若 t_x 表示真实死亡率，且 $w_x=n_x$，则 $w_x\cdot u_x$ 是 x 岁的观察死亡数，$w_x\cdot v_x$ 是修匀死亡数，那么 $\sum_x xw_x\cdot u_x$ 是观察总死亡年龄。类似地，$\sum_x xw_x\cdot v_x$ 是修匀总死亡年龄。F_1 较小就说明观察死亡人数接近修匀死亡人数，F_3 较小就说明总的观察死亡年龄接近总的修匀死亡年龄。F_1 或 F_3 越小，则死亡修匀值越能更真实地反映经验生命变化规律。

在由修匀值和初始值之间的偏差组成的序列中，还可根据其正负符号个数和模式确定拟合度量。这是一个相当简单但十分有用的方法。定义

$$d_x=v_x-u_x \tag{11.1.4}$$

从直观上可以看出，如果 d_x 的符号经常改变，意味着修匀序列与初始估计序列上下交错，而不是一方总在另一方的上方或下方。进一步地，如果估计量 U_x 是一个二项比例，那么观察（初始）估计值等可能地出现在 U_x 的上方或下方。又由于 U_x 的中位数与其均值 t_x 非常接近，如果 v_x 是 t_x 的最佳估计，则 u_x 出现在 v_x 的上方或下方的可能性大致相等。因此 d_x 出现正号或出现负号的可能性大致相等。

各个不同的 U_x 之间的独立性假设隐含着 d_j 与 d_{j-1} 的符号是相同的或相反的可能性是相等的。因此，从 d_{j-1} 到 d_j 符号改变的概率是 1/2。如果共有 n 个 d_j，则总序列 $\{d_j\}$ 符号改变的个数服从二项分布 $B(n-1,1/2)$，符号改变的个数的期望值为 $(n-1)/2$。所以如果修匀值与初始估计值拟合得很好时，则 $\{d_j\}$ 的符号改变频繁，接近 n/2 次。

11.1.3 修匀方法的分类

本章主要讨论两类修匀方法：假设真实的函数是 $t_x=f(x;c)$，其中 c 是未知参数，如果已知初始估计值序列 u_x，则

第一种方法是表格数据修匀。这种方法是直接对所获取的表格（或离散）数据进行修匀，即直接去找一个函数 G，用 $v_x=G(u_x)$ 作为 t_x 新的估计值。

第二种方法是参数修匀。这种方法是先去估计参数 c，再将 c 的估计值 $\hat{c}(u_x)$ 代入函数 f 中，用 $v_x=f(x;\hat{c}(u_x))$ 作为 t_x 新的估计值。所以参数修匀是间接的，而且必须要知道函数 f 的参数表达形式。

§11.2 表格数据修匀

11.2.1 移动加权平均修匀（M－W－A）

移动加权平均修匀（M－W－A）是发展较早的方法，它使用方便，在没有计算机的年代比较流行。M－W－A 认为修匀值是一组连续的、确定个数的未修匀值（观察者或初始估计）的一个加权平均。

1. 基本表达式。

$$v_x=\sum_{r=-n}^{n}a_r u_{x+r} \tag{11.2.1}$$

显然，每一个修匀值都是由 $2n+1$ 个连续初始估计值确定的。为方便起见，称 $2n+1$ 为该 M－W－A 的修匀幅度。

若 $a_r=a_{-r}$，$r=1,2,3,\cdots,n$，则称式（11.2.1）为对称 M－W－A 公式。

为了降低系数 a_r 的确定难度，本章仅限于讨论对称 M－W－A 公式。如无特别说明，本教材中提到的 M－W－A 公式都是对称的。

由于式（11.2.1）是中心对称（r 从 $-n$ 到 n），这便出现了“端值”问题，即如果 u_a 和 u_b 分别是具有最大和最小指标的 u_x，则根据式(11.2.1)，具有最大和最小指标的修匀值分别是 v_{a+n} 和 v_{b-n}，而不是 v_a 和 v_b。

2. 再生性。修匀的目的是使估计值尽可能地靠近其真实值。如果估计值碰巧是真实值，则一个好的修匀公式应使得修匀的结果仍然是真实估计值，即

$$t_x = \sum_{r=-n}^{n} a_r t_{x+r} \tag{11.2.2}$$

这里我们将这一性质称为修匀表达式的再生性。

实际的情况是真实值 t_x 并不知道，但是我们可能获得 t_x 的函数类型的某些先验信息，这时可通过式（11.2.2）来确定 a_r。

命题 11－1 若 t_x 是 x 三次多项式，则再生性条件式（11.2.2）等价于

$$\sum_{r=-n}^{n} a_r = 1, \sum_{r=-n}^{n} r^2 a_r = 0 \tag{11.2.3}$$

或者 $a_0 + 2\sum_{r=1}^{n} a_r = 1, \sum_{r=1}^{n} r^2 a_r = 0$ （11.2.4）

证明：设 $t_x = c_0 + c_1 x + c_2 x^2 + c_3 x^3$ （$c_3 \neq 0$），则式（11.2.2）可表示为：

$$\begin{aligned} & c_0 + c_1 x + c_2 x^2 + c_3 x^3 \\ & = c_0 \sum_{r=-n}^{n} a_r + c_1 \sum_{r=-n}^{n} a_r (x+r) + c_2 \sum_{r=-n}^{n} a_r (x+r)^2 + c_3 \sum_{r=-n}^{n} a_r (x+r)^3 \\ & = x^3 \cdot c_3 \left(\sum_{r=-n}^{n} a_r\right) + x^2 \cdot c_2 \left(\sum_{r=-n}^{n} a_r\right) + x\left[c_1 \left(\sum_{r=-n}^{n} a_r\right) + 3c_3 \sum_{r=-n}^{n} r^2 a_r\right] + c_0 \sum_{r=-n}^{n} a_r + c_2 \sum_{r=-n}^{n} r^2 a_r \end{aligned}$$

(其中第二个等式用了对称性)

比较三次项的系数知 $\sum_{r=-n}^{n} a_r = 1$ 。比较一次项的系数，并注意到 $c_3 \neq 0$，知 $\sum_{-n}^{n} r^2 a_r = 0$ 故式（11.2.3）成立。由对称性可知式（11.2.4）成立。

很多物理曲线都符合或可近似为多项式 $c_0 + c_1 x + c_2 x^2 + c_3 x^3$，所以实际中 t_x 常被假设符合三次多项式，因此式（11.2.3）也常作为再生性条件的等价形式。

3. 误差分析。

设 $U_x = t_x + E_x$，令 $V_x = G(U_x)$ 为修匀变量，$E_x' = G(E_x)$，则

$$V_x = G(t_x) + E_x' = \sum_{r=-n}^{n} a_r t_{x+r} + E_x' = \sum_{r=-n}^{n} a_r (t_{x+r} + E_{x+r}) \tag{11.2.5}$$

根据再生性，若 $E(E_x)=0$，则 $E(V_x)=t_x$，$E(E_x^{'})=0$。这说明如果初始估计值 U_x 是 t_x 的无偏估计，则修匀量 V_x 也是 t_x 的无偏估计。那么，V_x 作为 t_x 的估计量要优于 U_x，究竟体现在何处？这便是 V_x 较 U_x 更"稳定"。这种稳定性体现之一就是 $Var(V_x)$ 小于 $Var(U_x)$。

或者说，$R_0^2=\dfrac{Var(V_x)}{Var(U_x)}$的值越小，$V_x$ 较 U_x 就越优。

若假设 $\{E_{x+r}\}$ 相互独立且方差均为 σ^2，则

$$R_0^2=\frac{Var(V_x)}{Var(U_x)}=\frac{Var(V_x)}{Var(E_x)}=\frac{Var(\sum_{r=-n}^{n}a_rU_{x+r})}{\sigma^2}=\sum_{r=-n}^{n}a_r^2 \tag{11.2.6}$$

在公式（11.2.3）的条件下，$\min R_0^2$ 便可确定 a_r。此时得到 M－W－A 公式为使 R_0 最小化公式。$\gamma=(R_0^2)^{-1}$称为 M－W－A 公式的权重。

进一步，如果假设 t_x 是 x 三次多项式，再生性条件为

$$\sum_{r=-n}^{n}a_r=1,\sum_{r=-n}^{n}r^2a_r=0 \tag{11.2.7}$$

由最优化理论知，拉格朗日乘子可表示为：

$$L=\sum_{r=-n}^{n}a_r^2-\lambda(\sum_{r=-n}^{n}a_r-1)-\mu\sum_{r=-n}^{n}r^2a_r \tag{11.2.8}$$

所以可得$\dfrac{\partial L}{\partial a_r}=2a_r-\lambda-\mu r^2=0$，即有

$$a_r=\frac{1}{2}\lambda+\frac{1}{2}\mu r^2 \tag{11.2.9}$$

将式（11.2.9）代入式（11.2.3），得：

$$\begin{cases}a(2n+1)+2b\sum_{r=1}^{n}r^2=1\\ a\sum_{r=1}^{n}r^2+b\sum_{r=1}^{n}r^4=0\end{cases}$$

其中 $a=\dfrac{\lambda}{2}$，$b=\dfrac{\mu}{2}$，求解此方程组并代入式（11.2.9），得：

$$a_r=\frac{3(3n^2+3n-1)-15r^2}{(2n-1)(2n+1)(2n+3)}$$ ■

11.2.2 Whittaker 修匀

Whittaker 修匀于 1923 年开始形成和发展，其基本思想源于对修匀结果的光滑与拟合要求。

1. 基本表达式。Whittaker 修匀由两部分组成，一部分是拟合度量算子，另一部分是光滑度量算子。表达式如下：

$$M=F+hS=\sum_{x=1}^{n}w_x(v_x-u_x)^2+h\sum_{x=1}^{n-z}[\Delta^zv_x]^2 \tag{11.2.10}$$

上式中拟合度量算子 $F = \sum_{x=1}^{n} w_x (v_x - u_x)^2$ 类似最小二乘法的思想。w_x 为权重，往往确定为$\frac{n_x}{v_x (1 - v_x)}$或$\frac{n_x}{\bar{n}}$，其中 $\bar{n}$ 是 n_x 关于所有 x 的算术平均。这主要是为了体现观察死亡人数与修匀死亡人数的差别。光滑度量算子是 $S = \sum_{x=1}^{n-z} [\Delta^Z v_x]^2$，其中 Δ^Z 为 Z 阶差分算子。曲线的光滑性在连续时是通过导数来衡量的，可导阶数越高曲线越光滑，而差分运算和求导数是对应的，所以差分的阶数越高对光滑性要求也就越高。参数 h 是一个权重，它在拟合和光滑性之间作了一个平衡。

值得注意的是，在式（11.2.10）中，下标 $x=1$ 表示第一个最小年龄，$x=n$ 表示最后一个最大年龄。最小、最大年龄到底是多少？这取决于所考虑的问题本身。

通过最小化 M，可确定修匀值 v_x。这实际上是一个无约束的二次优化问题。可以证明，使式（11.2.10）极小化的修匀值 v_x 必自动地使拟合度量 $F_1 = \sum_x w_x(v_x - u_x)$，$F_3 = \sum_x x w_x(v_x - u_x)$ 为零。

因此，对死亡数据而言，取 $w_x = n_x$ 时，有

$$\sum_x n_x v_x = \sum_x n_x u_x, \sum_x x n_x v_x = \sum_x x n_x u_x$$

这说明，对观察数据 u_x 和修匀数据 v_x 来说，死亡总人数是相等的，死亡总年龄也是相等的。

2. M 的极小化。显然，M 是 n 个未知值 v_x 的一个函数，因而，极小化 M 的 v_x 是以下 n 个方程的解：

$$\frac{\partial M}{\partial v_r} = 0, \quad r = 1, 2, \cdots, n \tag{11.2.11}$$

虽然用这种标准计算技巧可以求出 M 的极小值，但是我们这里介绍另一种方法解决极小化问题，那就是应用关于 M 的一种矩阵—向量公式。用 $\mathbf{y}'$表示矩阵 $\mathbf{y}$ 的转置。令

$$\mathbf{u}' = [u_1, u_2, \cdots, u_n], \quad \mathbf{v}' = [v_1, v_2, \cdots, v_n],$$

$$\mathbf{w} = \begin{bmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{bmatrix}$$

则 $F = (\mathbf{v} - \mathbf{u})' \mathbf{w} (\mathbf{v} - \mathbf{u})$。

为了用矩阵表示 hS，考虑展式

$$(1 - x)^z = 1 - C_z^1 x + C_z^2 x^2 + \cdots + (-1)^k C_z^k x^k + \cdots + (-1)^z x^z \tag{11.2.12}$$

矩阵 $\mathbf{K}_z$ 为$(n-z) \times n$ 矩阵，其元素按照下面的方式确定：

（1）若 z 为偶数，$\mathbf{K}_z$ 的第一行中前 $z+1$ 个元素恰为式（11.2.12）中的 $z+1$ 个系数，后 $n-z+1$ 个元素为 0。第二行元素是将第一行右移一格得到。被挤出的一个 0 放到前面去。用类似的方法可得其余各行元素。

（2）若 z 为奇数，则 $\mathbf{K}_z$ 的第一行中前 $z+1$ 个元素恰为式（11.2.12）中的 $z+1$ 个系数的相反数，后 $n-z+1$ 个元素为 0，其余各行元素仍按上述右移方法得到。

（3）$hS=h(\mathbf{K}_z\mathbf{v})'(\mathbf{K}_z\mathbf{v})$。

【例 11－1】 分别在下列情形下，求矩阵 $\mathbf{K}$ 及 S 的形式：（1）$z=2$，$n=6$；（2）$z=3$，$n=7$。

解：（1）2 阶差分算子可以通过矩阵表达为：

$$\mathbf{K}_2=\begin{bmatrix}1 & -2 & 1 & 0 & 0 & 0\\ 0 & 1 & -2 & 1 & 0 & 0\\ 0 & 0 & 1 & -2 & 1 & 0\\ 0 & 0 & 0 & 1 & -2 & 1\end{bmatrix}$$

$$\mathbf{K}_2\mathbf{v}=[\Delta^2 v_1,\ \Delta^2 v_2,\ \Delta^2 v_3,\ \Delta^2 v_4]$$

$$S=(\mathbf{K}_2\mathbf{v})'(\mathbf{K}_2\mathbf{v})=\mathbf{v}'(\mathbf{K}'_2\mathbf{K}_2)\mathbf{v}=\sum_{x=1}^{4}[\Delta^2 v_x]^2$$

（2）$$\mathbf{K}_3=\begin{bmatrix}-1 & 3 & -3 & 1 & 0 & 0 & 0\\ 0 & -1 & 3 & -3 & 1 & 0 & 0\\ 0 & 0 & -1 & 3 & -3 & 1 & 0\\ 0 & 0 & 0 & -1 & 3 & -3 & 1\end{bmatrix}$$

可知 $S'=\sum_{x=1}^{4}(\Delta^3 v_x)^2=\mathbf{v}'(\mathbf{K}'_3\mathbf{K}_3)\mathbf{v}$。 ■

不难验证，如此定义的 $\mathbf{K}_z$ 为 n－z 维列向量，其元素恰为 $\Delta^z v_1$，$\Delta^z v_2$，$\cdots,\Delta^z v_{n-z}$。

最后，显然有 M 的如下矩阵形式：

$$\mathbf{M}=(\mathbf{v}-\mathbf{u})'\mathbf{w}(\mathbf{v}-\mathbf{u})+h\mathbf{v}'(\mathbf{K}'_z\mathbf{K}_z)\mathbf{v} \tag{11.2.13}$$

根据式（11.2.11），极小化 M。令 $\frac{\partial M}{\partial \mathbf{v}}=\left[\frac{\partial M}{\partial v_1},\cdots,\frac{\partial M}{\partial v_n}\right]'$，由 $\frac{\partial M}{\partial \mathbf{v}}=0$，得线性方程组

$$[\mathbf{w}+h\mathbf{K}'_z\mathbf{K}_z]\mathbf{v}=\mathbf{wu}\text{ 或 }\mathbf{cv}=\mathbf{wu} \tag{11.2.14}$$

这里，$\mathbf{c}=\mathbf{w}+\mathbf{h}\mathbf{K}'_z\mathbf{K}_z$。若 $\mathbf{c}$ 为非奇异矩阵，则

$$\mathbf{v}=\mathbf{c}^{-1}\mathbf{wu}$$

这一部分实际上只是用矩阵的算法求解优化问题（11.2.10）。

11.2.3　二维 Whittaker 修匀

前面所讨论的修匀都是一维序列的修匀，修匀序列的下标是一维的，

如死亡年龄。但在精算实务中，往往碰到待修匀数据的下标是二维的。例如，在选择—终极寿命表中，一个下标变量是选择年龄（投保年龄），用 $[x]$ 表示，另一个下标变量是从选择年龄开始的保障期，用 r 表示。因此 $q_{[x]+r}$表示选择年龄为 x 岁、满 $x+r$ 岁后一年内死亡的概率。沿用本章的记号，用 $U_{[x]+r}$表示初始估计量，$t_{[x]+r}=q_{[x]+r}$表示真实死亡率，$n_{[x]+r}$表示暴露数，$v_{[x]+r}$表示修匀值。初始的修正受着关于阵列 $t_{[x]+r}$的先验观点的指导，而且这些先验观点还包括着光滑性要求，同时要设法保持对于观察数据的某种拟合度。

在选择—终极寿命表中，下标的二维性不仅仅表现在选择年龄，还有死亡年龄对死亡率产生的影响。例如 $t_{[35]}<t_{[33]+2}$，也就是说，投保年龄为 35 岁的人一年内的死亡率比一个 35 岁但已投保两年的保户一年内的死亡率小。一般地，较小的 m 有 $t_{[x-m]+m}<t_{[x-m-1]+m+1}$。但是，实践表明，随着时间推移，选择年龄对死亡率的效用就减少了，最后完全取决于死亡年龄。选择年龄对死亡率产生作用的年限叫做**选择期**，超过选择期后的死亡率叫做**终极死亡率**。如若选择期为 k，则 $t_{[x]+r}=t_{x+r}$，$(r\geqslant k)$。

为了方便起见，选择—终极表用表 11－1 说明，其中用的是我们要估计的真实死亡率。我们任意选择一个选择期限 $k=4$。

表 11－1　　　　选择—终极表形式

选择年龄［x］	死亡率				终极死亡率	到达年龄
	选择经过的时期					
	1	2	3	4		
30	$t_{[30]}$	$t_{[30]+1}$	$t_{[30]+2}$	$t_{[30]+3}$	t_{34}	34
31	$t_{[31]}$	$t_{[31]+1}$	$t_{[31]+2}$	$t_{[31]+3}$	t_{35}	35
32	$t_{[32]}$	$t_{[32]+1}$	$t_{[32]+2}$	$t_{[32]+3}$	t_{36}	36
33	$t_{[33]}$	$t_{[33]+1}$	$t_{[33]+2}$	$t_{[33]+3}$	t_{37}	37
⋮	⋮	⋮	⋮	⋮	⋮	⋮

关于这组数据的先验观点除了在每一横行和每一竖列满足光滑性外，还有单调性要求，即关于死亡率的递增或递减模式的先验观点。对于保险数据，在每一列中，死亡率是递增的。因为随着选择年龄的增加，而从选择期开始保障期限都是相同的，所以死亡率是递增的。同样，在每一行中，死亡率是递增的，因为年龄和自选择期的保障期都在增加。进一步，在以下这种对角线上，死亡率也是递增的。这些对角线是从矩阵的左下到右上，在同一条这种对角线上，每个人的到达年龄都是相同的，但自选择期开始的保障期限却是递增的。因而可以认为：

$$t_{[35]}<t_{[34]+1}<t_{[33]+2}<t_{[32]+3}<t_{[31]+4}<t_{35}$$

这些先验观点：光滑性和上述三个方向的死亡率的递增性，再加上拟合性都是我们在修匀过程中需要考虑的。

在上面的叙述中，死亡率随保障期的增加而增加，这是正选择的情形。与此相反，对于医疗临床数据，同一年龄的死亡率都是保障期的减函数，这是负选择的情形。例如，投保的原因是需要治疗的一种特殊的疾病或者要做某个外科手术。此时，在选择以后很短的时间内，期望死亡率是很高的，而只有在治疗成功而继续生存的情况下，死亡率将逐渐减少。在这种情况下，每一行中的死亡率应该是递减的，而且在左下到右上的对角线上，死亡率也是递减的。

由于数据是二维的，所以它相当于一个曲面。如果利用 Whirraker 修匀的思想，则需要定义垂直光滑度量和水平光滑度量，再连同拟合度量一起，组成它们的线性组合，再通过极小化这个组合度量找出修正估计阵列。

假设有 $m \times n$ 个初始估计值 $u_{ij}, i=1,2,\cdots,m; j=1,2,\cdots,n$。这里 m 为选择年龄，n－1 为选择期。于是每个 i，有 n－1 个选择死亡率，一个终极死亡率。此时的拟合度量为：

$$F = \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}(v_{ij} - u_{ij})^2 \tag{11.2.15}$$

光滑性分别从垂直与平行两个方向得到。设 $\underset{v}{\Delta}^z$ 为垂直方向差分算子，$\underset{h}{\Delta}^y$ 为水平方向差分算子，则第 j 列的垂直光滑度量为 $\sum_{i=1}^{m-z}(\underset{v}{\Delta}^z v_{ij})^2$，总的垂直光滑度量为

$${}^{v}S = \sum_{j=1}^{n}\sum_{i=1}^{m-z}(\underset{v}{\Delta}^z v_{ij})^2 \tag{11.2.16}$$

类似地，总的水平光滑度量为：

$${}^{h}S = \sum_{i=1}^{m}\sum_{j=1}^{n-y}(\underset{h}{\Delta}^y v_{ij})^2 \tag{11.2.17}$$

于是，Whirraker 修匀算子为：

$$M = F + \alpha \cdot {}^{v}S + \beta \cdot {}^{h}S \tag{11.2.18}$$

为了完成 M 的极小化，把 M 写成矩阵形式。

1. F 的矩阵形式。将 $(u_{ij})_{m\times n}$ 排成 $mn \times 1$ 列向量，先排第 1 行，再排第 2 行，依次进行下去，则 u_{ij} 变成 $u_{n(i-1)+j}$。把这个列向量记为 $\mathbf{u}$，修匀向量记为 $\mathbf{v}$。

将权矩阵 $(w_{ij})_{m\times n}$ 变成 $mn \times mn$ 对角阵。方法是将权矩阵各行元素逐一放在主对角线上，则 w_{ij} 变成 $w_{n(i-1)+j,n(i-1)+j}$，把这个对角阵记为 $\mathbf{w}$。

F 的矩阵形式可表示为：

$$F = (\mathbf{v} - \mathbf{u})'\mathbf{w}(\mathbf{v} - \mathbf{u}) \tag{11.2.19}$$

2. ${}^{h}S$ 的矩阵形式。定义 $m(n-y) \times mn$ 矩阵，记为 ${}^{h}\mathbf{K}_y$，使得 ${}^{h}\mathbf{K}_y\mathbf{v}$ 包含

的$m(n-y)$个元素$\underset{h}{\Delta}^{y} v_{ij}$的次序正好与式（11.2.17）相同，则

$$^{h}S = ({}^{h}\mathbf{K}_{y}\mathbf{v})'\ ({}^{h}\mathbf{K}_{y}\mathbf{v}) \tag{11.2.20}$$

矩阵$^{h}\mathbf{K}_{y}$可按如下方法得到：

（1）按照一维 Whirraker 修匀的方法，将 $\mathbf{v}$ 换成矩阵 $\mathbf{U}=(u_{ij})_{m\times n}$的第 1 行元素，得到的矩阵记为$_{1}^{h}\mathbf{K}_{y}$；类似地，再换成 $\mathbf{U}$ 的第 2 行元素，得到的矩阵记为$_{2}^{h}\mathbf{K}_{y}$，依次继续下去，最后$_{m}^{h}\mathbf{K}_{y}$。它们都是$(n-y)\times n$ 矩阵。

（2）令

$$^{h}\mathbf{K}_{y} = \begin{bmatrix} _{1}^{h}\mathbf{K}_{y} & & & \\ & _{2}^{h}\mathbf{K}_{y} & & \\ & & \ddots & \\ & & & _{m}^{h}\mathbf{K}_{y} \end{bmatrix}$$

容易验证它符合上述要求。

3. ^{v}S 的矩阵形式。定义 $n(m-z)\times mn$ 矩阵$^{v}\mathbf{K}_{z}$，使得

$$^{v}S = ({}^{v}\mathbf{K}_{z}\mathbf{v})'({}^{v}\mathbf{K}_{z}\mathbf{v})$$

极小化 M 后，得方程组

$$(\mathbf{w} + \alpha\,{}^{v}\mathbf{K}'_{z}\,{}^{v}\mathbf{K}_{z} + \beta\,{}^{h}\mathbf{K}'_{y}\,{}^{h}\mathbf{K}_{y})\,\mathbf{v} = \mathbf{wu}$$

【例 11－2】 已知观察值 $\mathbf{U}$ 和修匀值 $\mathbf{V}$ 分别如下：

$$\mathbf{U} = \begin{bmatrix} 1 & 1 & 6 \\ 3 & 4 & 6 \end{bmatrix},\quad \mathbf{V} = \begin{bmatrix} 2 & 3 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

并且知道：

（1）所有的权数都等于 2；

（2）$F = \sum_{i=1}^{2}\sum_{j=1}^{3} w_{ij}(u_{ij}-v_{ij})^{2}$；

（3）垂直光滑度量^{v}S 和水平光滑度量^{h}S 都采用一阶差分；

（4）$M = F + 2\,{}^{v}S + {}^{h}S$。

求 M。

解：

$$^{v}S = \sum_{j=1}^{3}\sum_{i=1}^{2}(\Delta v_{ij})^{2} = \sum_{j=1}^{3}(\Delta v_{ij})^{2} = (3-2)^{2} + (5-3)^{2} + (6-5)^{2} = 6$$

$$^{h}S = \sum_{i=1}^{2}\sum_{j=1}^{2}(\Delta v_{ij})^{2} = (3-2)^{2} + (5-3)^{2} + (5-3)^{2} + (6-5)^{2} = 10$$

$$F = \sum_{j=1}^{2}\sum_{i=1}^{3} 2(u_{ij}-v_{ij})^{2} = 14$$

则 $\quad M = 14 + 12 + 10 = 36$

11.2.4 Baysian 修匀

Baysian 修匀与数理统计中的 Baysian 估计十分相似。下面将对这种方

法进行简单的描述。

就像 Baysian 估计一样，Baysian 修匀按如下四个步骤进行：

（1）确定 t_x 的先验分布。设 T 为一个连续型向量，t_x 是随机变量 T 的一个观察值。其先验密度为 $f_T(t)$。

（2）确定实验模型。所谓实验模型，就是估计量 $\mathbf{U}$ 的条件密度，记为 $f_{U|T}(\mathbf{u}|\mathbf{t})$。

（3）确定后验分布。这就是在已知样本的前提下 T 的分布：

$$f_{T|U}(\mathbf{u}|\mathbf{t})=\frac{f_{U|T}(\mathbf{t}|\mathbf{u})\cdot f_T(\mathbf{t})}{f_U(\mathbf{u})} \tag{11.2.21}$$

（4）修匀值 v_x 的确定。v_x 是 t_x 的最佳估计，若这种最佳的衡量标准是均方误差最小的话，则该值应为 V 关于 U 条件期望，即 $\mathbf{V}=E(\mathbf{T}|\mathbf{U})$。

【例 11－3】 将一枚硬币上抛 n 次，记其正面出现的次数为 H，试用 Baysian 修匀，确定其正面出现的概率 t［已知 t 的先验分布为 β（a，b）］。

解：（1）先验分布。

$$f_T(t)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}t^{a-1}(1-t)^{b-1},\quad 0\leqslant t\leqslant 1,\ a>0,\ b>0 \text{ 为参数}$$

（2）实验模型。H 的分布律为：

$$P_{H|T}(h|t)=C_n^h t^h(1-t)^{n-h},\qquad h=0,1,2,\cdots,n$$

H 的边缘分布律为：

$$P_H(h)=\int_0^1 P_{H|T}(h|t)f_T(t)dt=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot C_n^h\cdot\frac{\Gamma(a+h)\Gamma(b+n-h)}{\Gamma(a+b+n)}$$

（3）后验分布。T 的后验密度为：

$$f_{T|H}(t|h)=\frac{P_{H|T}(h|t)\cdot f_T(t)}{P_H(h)}=\frac{\Gamma(a+b+n)}{\Gamma(a+h)\Gamma(b+n-h)}\cdot t^{a+h-1}(1-t)^{b+n-h-1}$$

即为贝塔分布 $\beta(a+h,\ b+n-h)$。

（4）确定修匀值。

$$v=E(T|H)=\frac{a+h}{a+b+n}=\frac{a}{a+b}\cdot\frac{a+b}{a+b+n}+\frac{h}{n}\cdot\frac{n}{a+b+n} \tag{11.2.22}$$

式（11.2.22）说明修匀值正是先验估计值$\frac{a}{a+b}$与后验估计值$\frac{h}{n}$的加权平均，权数分别是$\frac{a+b}{a+b+n}$及$\frac{n}{a+b+n}$。■

下面介绍两种常见的 Baysian 修匀。

1. Kimeldorf－Jones 方法。该方法是一种特殊的 Baysian 修匀，该方法其实就是在先验分布为多维正态分布情形下的 Baysian 修匀。

（1）T 的先验分布。若先验分布是一维正态分布，则其密度函数可表示为：

$$f_T(t)=(2\pi a)^{-\frac{1}{2}}\cdot\exp\left[-\frac{1}{2}a^{-1}(t-m)^2\right]$$

这里，m 是均值，a 为其方差。

若先验分布为 n 维正态分布，则其联合分布密度函数为：

$$f_{\mathbf{T}}(\mathbf{t})=[(2\pi)^{n}|\mathbf{A}|]^{-\frac{1}{2}}\cdot\exp\left[-\frac{1}{2}(\mathbf{t}-\mathbf{m})'\mathbf{A}^{-1}(\mathbf{t}-\mathbf{m})\right]$$

这里 $\mathbf{t}=(t_1,\cdots,t_n)'$，$\mathbf{m}=(m_1,\cdots,m_n)'$为其均值向量，$\mathbf{A}$ 为其协方差矩阵，且 $\mathbf{A}$ 是对称正定矩阵。

（2）实验模型。在 T 已知的情况下，U 的条件密度可表示为：

$$f_{U|T}(\mathbf{u}\mid\mathbf{t})=[(2\pi)^{n}|\mathbf{B}|]^{-\frac{1}{2}}\cdot\exp\left[-\frac{1}{2}(\mathbf{u}-\mathbf{t})'\mathbf{B}^{-1}(\mathbf{u}-\mathbf{t})\right]$$

这里，$u=(u_1,\cdots,u_n)'$，B 为协方差矩阵，且是对称正定矩阵。

（3）后验分布。在 U 已知的情况下，T 的后验密度为：

$$f_{T|U}(\mathbf{t}\mid\mathbf{u})=k_1\exp\left\{-\frac{1}{2}[(\mathbf{t}-\mathbf{m})'\mathbf{A}^{-1}(\mathbf{t}-\mathbf{m})+(\mathbf{u}-\mathbf{t})'\mathbf{B}^{-1}(\mathbf{u}-\mathbf{t})]\right\}$$

$$=k_2\exp\left[-\frac{1}{2}(\mathbf{t}-\mathbf{v})'\mathbf{C}^{-1}(\mathbf{t}-\mathbf{v})\right]$$

$$\mathbf{v}=(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{B}^{-1}\mathbf{u}+\mathbf{A}^{-1}\mathbf{m}) \tag{11.2.23}$$

$$\mathbf{C}=(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}$$

这里 k_1，k_2 分别是不含 t 的确定系数。

（4）确定修匀值。$f_{T|U}$（$\mathbf{t}\mid\mathbf{u}$）的均值、众数或中位数，都可取做修匀向量。在 K－J 方法中，由于后验分布是多元正态分布，这三者是完全相同的。这个共同的向量，我们已用 $\mathbf{v}$ 表示修匀向量，由式（11.2.23）给出。值得注意的是，由于式（11.2.23）中需要对两个较难的矩阵求逆，因此将式（11.2.23）重新整理为：

$$\mathbf{v}=\mathbf{u}+(\mathbf{I}+\mathbf{AB}^{-1})^{-1}(\mathbf{m}-\mathbf{u})=\mathbf{m}+(\mathbf{I}+\mathbf{BA}^{-1})^{-1}(\mathbf{u}-\mathbf{m}) \tag{11.2.24}$$

这里 $\mathbf{I}$ 是 n 阶单位阵。$\mathbf{B}$ 是对角阵，容易求逆。所以式（11.2.4）只需要对一个较难的矩阵求逆。

从式（11.2.23）和（11.2.24）可以看出，元素 m、A 和 B 都是修匀方法的参数，为了实施修匀，必须要确定它们。London 讨论了参数确定指导性原则，本书在此不叙述①。

2. Dirichlet 修匀。前面所讨论的修匀，着重于估计不同年龄的死亡率。然而，在有些参数生存模型中，年龄并不是重要因素，可以不考虑，而只考虑不同时间区间上的死亡率。如一种特殊手术后病人的生存概率。这类问题，可以用 Bayes 修匀的思想进行修匀。下面介绍的 Dirichlet 修匀就是处理这类问题的一种方法。

设 D_i 为第 i 个时间区间的死亡人数，开始观察时，群体人数为 d，最

① 参见 Dick London 著、徐诚浩译：《修匀数学》，上海科学技术出版社 1996 年版，第 98～102 页。

后一个死亡最迟发生在第 n 个时间区间内，则显然有 $\sum_{i=1}^{n} D_i = d$。令 $\mathbf{D} = (D_1, \cdots, D_n)$，则 $\mathbf{D}$ 的分布是多项分布 $\mathbf{B}(d_1, \cdots, d_n)$，其分布为：

$$P_{D|T}(\mathbf{d} \mid \mathbf{t}) = \frac{d!}{\prod_{i=1}^{n} d_i!} \prod_{i=1}^{n} t_i^{d_i}, 0 \leqslant t_i \leqslant 1, i = 1, \cdots, n \qquad (11.2.25)$$

对于式（11.2.25）有以下几点说明：

（1）$\mathbf{d} = (d_1, \cdots, d_n)'$是不同时间区间的死亡人数构成的列向量；

（2）$\mathbf{t} = (t_1, \cdots, t_n)'$，$t_i$ 为某个人在第 i 个时间区间上的死亡率，用精算符号表示为${}_{i-1|}q_0$；

（3）$\sum_{i=1}^{n} d_i = d, \sum_{i=1}^{n} t_i = 1$。

令 $\mathbf{T} = (T_1, \cdots, T_n)'$，$T_i$ 是 Bayes 观点中 t_i 所对应的随机变量。在 Dirichlet 修匀中，假设 T 的先验分布是 Dirichlet 分布，其密度函数为：

$$f_{\mathbf{T}}(\mathbf{t}) = \frac{\Gamma(\sum_{i=1}^{n} a_i)}{\prod_{i=1}^{n} \Gamma(a_i)} \prod_{i=1}^{n} t_i^{a_i - 1}, a_i > 0, 0 \leqslant t_i \leqslant 1, i = 1, \cdots, n \qquad (11.2.26)$$

令 $a = \sum_{i=1}^{n} a_i$，则易知 $E(T_i) = a_i / a$。

经计算，T 的后验分布仍是 Dirichlet 分布，密度函数为：

$$f_{\mathbf{T|D}}(\mathbf{t} \mid \mathbf{d}) = \frac{\Gamma(\sum_{i=1}^{n} (a_i + d_i))}{\prod_{i=1}^{n} \Gamma(a_i + d_i)} \prod_{i=1}^{n} t_i^{(a_i + d_i) - 1}, a_i > 0, 0 \leqslant t_i \leqslant 1, i = 1, \cdots, n$$

由 Baysian 修匀的结论知：

$$v_i = E(T_i \mid d_i) = \frac{a_i + d_i}{a + d}$$

用矩阵表示为 $\mathbf{v} = \frac{1}{a + d}(\mathbf{a} + \mathbf{d})$，其中 $\mathbf{a} = (a_1, \cdots, a_n)$。

【例 11－4】 考察 200 只被注射某种药物的老鼠未来每年的存活情况，已知这些老鼠最多存活 10 年，且先验观点是 $\mathbf{T} = (T_1, \cdots, T_{10})'$服从 Dirichlet 分布，参数 $\mathbf{a} = (a_1, \cdots, a_{10})$ 满足 $\sum_{i=1}^{10} a_i = 600, a_1 = 15, \{a_i\}_{i=1}^{10}$ 是一个等差级数。试求这群老鼠的生存概率。

解： 由题意 $a_i = a_1 + (i-1)k$，$i = 1, \cdots, 10$，因此 $600 = \sum_{i=1}^{10} a_i = 150 + 45k$，从而推出 $k = 10$。因而，

$$v_i = {}_{i-1|}q_0 = \frac{a_i + d_i}{a + d} = \frac{25 + 10i}{800} = \frac{5 + 2i}{160}$$

所以，

$$S(i-1)-S(i)=\frac{5+2i}{160} \qquad i=1,\cdots,10$$

于是有

$$S(i)=1-\frac{i^2+6i}{160} \qquad i=1,\cdots,10$$ ■

表格数据修匀方法很多，如图表修匀、方差补整修匀和参照标准表修匀等，它们都比较直观易懂，这里就不一一介绍了，有兴趣的读者可以查阅相关文献。

§11.3 参数修匀

前面所介绍的表格数据修匀都是取定一个观察序列 u_x，作为对真实序列 t_x 的初始估计序列，再根据关于 t_x 的先验观点，修改这个序列的值，产生一个修正估计序列 v_x。修正序列 v_x 都表达成数值形式。在本节中，我们所讨论的参数修匀法与表格数据修匀法的基本差别在于，修正序列表达成自变量 x 的一个数学函数，因而修匀值记号 v_x 就表示成 x 的带参函数，这些参数必须由观察数据所确定。当然，所得函数能用来对所有所需的 x 求值，如果需要的话，修匀值还能以表格形式表示。

11.3.1 函数形式

在参数修匀法中，估计问题已归结为曲线拟合问题。本节将讨论的三种函数形式都是用来表示死亡数据的，是关于死亡力 μ_x 的含参函数形式。根据寿险精算理论，死亡率 q_x 与死亡力 μ_x 的关系式是：

$$q_x = 1 - e^{-\int_0^1 \mu_{x+t} dt}$$

因此，从本质上看，μ_x 和 q_x 一一对应。

由于在第二章中，对几种特殊的死亡力形式已作过详细讨论。因此，这里只做些简单的描述。

1. Gompertz 形式。

$$\mu_x = Bc^x, \qquad B>0,\ c>1$$

两边对数变换：

$$\log\mu_x = \log B + (\log c)x \tag{11.3.1}$$

一个具体的问题是：所给的死亡力数据是否适合这个模型？解决这个问题的方法是数据合适性预检。如果数据适合 Gompertz 形式，则 $\log\mu_x$ 相对于 x 的图像应近似为一条直线；如果得不到这种近似直线，那么应考虑另一种函数形式。

如果初始估计 u_x 是生存概率，此时 Gompertz 形式是：

$$p_x = g^{c^x(c-1)}$$

这里，g，c 均是确定性常数。对数变换为

$$\log p_x = c^x(c-1)(\log g)$$

于是，

$$\frac{\log p_{x+1}}{\log p_x} = c$$

因此，数据合适性预检方法是看 $\log u_{x+1}/\log u_x$ 是否接近于常数。

2. Weibull 形式。

$$\mu_x = k(x-a)^n,\ x \geqslant a,\ k>0,\ n>0$$

或当 $a=0$，$\mu_x = kx^n$。其对数形式为：

$$\log\mu_x = \log k + n\log x \tag{11.3.2}$$

数据合适性预检法是检查 $\log u_x$ 关于 $\log x$ 的曲线是否近似为一条直线。

如果初始估计 u_x 是生存概率，有

$$p_x = \exp\left\{-\frac{k}{n+1}\left[(x+1)^{n+1} - x^{n+1}\right]\right\}$$

和 $\ln p_x = -\frac{k}{n+1}\left[(x+1)^{n+1} - x^{n+1}\right] = -\frac{k}{n+1}\Delta x^{n+1}$

如果 n 是整数，则

$$\Delta^n \ln p_x = -\frac{k}{n+1}\Delta^{n+1}x^{n+1} = -\frac{k}{n+1}(n+1)! = -kn!$$

是常数。因此，当 u_x 是生存概率而非死亡力，数据的预检方法就是构造关于 $\ln u_x$ 的差分表，看它的 n 阶差分是否接近于常数。

3. Makeham 形式。

$$\mu_x = A + Bc^x \text{ 和 } p_x = s\cdot g^{c^x(c-1)}$$

在此情形下，$\log\mu_x$ 既不是线性，也不是对数—线性的，因此在最小二乘拟合时，必须要做些修正①。在数据 u_x 为死亡力和生存概率时的数据预检法是什么呢？留给读者作为练习。

11.3.2　参数估计

1. 配置法。参数估计的配置法是利用恰当的差分运算来确定参数值。为说明这种方法，假设 u_x 是 p_x 的初始估计值，下面采用 Makeham 形式来说明。其等价形式为：

$$p_x = sg^{c^x(c-1)} \tag{11.3.3}$$

选取三个等间隔的自变量 x，$x+r$，$x+2r$，有

① 参见 Dick London 著、徐诚浩译：《修匀数学》，上海科学技术出版社 1996 年版，第 122 页。

$$\frac{\Delta_r \log u_{x+r}}{\Delta_r \log u_x} = c^r$$

Δ_r 是间隔为 r 的向前差分，即

$$\Delta_r f(x) = f(x+r) - f(x)$$

由此可确定 c。由 $\Delta_r \log u_x = (c-1)(c^r-1)(\log g)c^x$ 确定 g。最后由

$$\Delta_r \log u_x = \log s + (c+1)(\log g)c^x$$

可确定 s。

事实上，这个模型中有三个参数，已知函数中的三个点相当于确定了三个方程，在一定条件下，可以解出所有的参数。同样如果初始估计值多于三个，那么根据配置法，就可能得到参数的不同估计，此时，取舍问题非常棘手，这便是这种方法的不足之处。

2. 最小二乘修匀法。考虑一般的线性关系式：

$$u_x^* = a + bx^* \tag{11.3.4}$$

注意式（11.3.1）就是这种形式，其中 $u_x^* = \log u_x$ 和 $x^* = x$。类似地，式（11.3.2）也是这种形式。对式（11.3.4）构造最小二乘函数：

$$SS = \sum_{x=1}^{n} w_x (u_x^* - a - bx^*)^2$$

极小化 SS，令 $\frac{\partial SS}{\partial a} = 0$，$\frac{\partial SS}{\partial b} = 0$ 得到线性方程组，求解可得 a 和 b 的值。

若 u_x 是非线性的，可通过变换使其线性化。如 Compertz 模型 $\mu_x = Bc^x$，取对数则有

$$\log \mu_x = \log B + x \log c$$

若已知 μ_x 的观测值序列 $\{u_x\}$，则用最小二乘法估计参数 $\log B$，$\log c$ 之后，任给 x，都可以得到一个 μ_x 的修匀值或是新的估计值。

3. 极大似然法。在精算实务中，若死亡力的含参函数形式已知，则各种事件（如死亡、退出等）的概率值的形式也已知，于是可利用极大似然估计法获取参数的估计，但样本信息不同，似然函数也就不同。下面分情况讨论这种方法。

（1）完整数据、精确死亡年龄。此时的样本信息是：已知第 i 个人进入观察的年龄为 x_i，死亡年龄为 y_i（$i = 1, \cdots, n$），于是对于每个人（事件）所对应的似然因子为：

$$L_i = {}_{y_i - x_i}p_{x_i} \mu_{y_i} = \exp\left(-\int_{x_i}^{y_i} \mu_s ds\right)\mu_{y_i} \tag{11.3.5}$$

总体似然函数为：

$$L = \prod_{i=1}^{n} L_i$$

则

$$\ln L = \sum_{i=1}^{n}(\ln\mu_{y_i} - \int_{x_i}^{y_i}\mu_s ds) \tag{11.3.6}$$

若已知 μ_x 的参数形式，通过极大化 $\ln L$，可得参数估计。

（2）非完整数据、精确死亡年龄或退出年龄。此时的样本信息为：在观察期内，每个人要么死亡，要么退出。其中，第 i 个人，若死亡，死亡年龄记为 y_i；若退出，所观察到的退出年龄记为 $z_i(i=1,\cdots,n)$。死亡者的全体记为 D，退出者的全体记为 W。此时，每个在 z_i 退出的人对应的似然因子为：

$$L_i = {}_{z_i-x_i}p_{x_i} = \exp(-\int_{x_i}^{z_i}\mu_s ds) \tag{11.3.7}$$

这里，x_i 为进入观察的年龄，下同。

每个死亡人的似然因子为式（11.3.5），统一式（11.3.5）与式（11.3.7）的形式，无论如何，第 i 个人的似然因子可表示为：

$$L_i = \exp(-\int_{x_i}^{w_i}\mu_s ds)(\mu_{w_i})^{a_i} \tag{11.3.8}$$

其中，

$$a_i = \begin{cases} 0, & \text{第 } i \text{ 个人活着（退出）} \\ 1, & \text{第 } i \text{ 个人死亡} \end{cases}, \quad w_i = \begin{cases} z_i, & \text{第 } i \text{ 个人活着（退出）} \\ y_i, & \text{第 } i \text{ 个人死亡} \end{cases}$$

总体似然函数可表示为：

$$L = \prod_{i\in D}\mu_{y_i}\exp(-\int_{x_i}^{y_i}\mu_s ds)\prod_{i\in W}\exp(-\int_{x_i}^{z_i}\mu_s ds) = \prod_{i=1}^{n}\exp(-\int_{x_i}^{w_i}\mu_s ds)(\mu_{w_i})^{a_i}$$

则

$$\ln L = \sum_{i=1}^{n}(a_i\ln\mu_{w_i} - \int_{x_i}^{w_i}\mu_s ds) \tag{11.3.9}$$

若 μ_x 的参数形式已知，则通过极大化 $\ln L$，可得参数估计。

【例 11-5】 某一死亡力研究从 1984 年 1 月 1 日开始，到 1988 年 1 月 1 日结束，观察结果如表 11-2 所示。

死亡力 $\mu_x = kx$，求 k 的极大似然估计。

表 11-2

	1984 年 1 月 1 日的年龄	观察结果
甲	1	1986. 1. 1 死亡
乙	2	1988. 1. 1 仍生存
丙	3	1986. 1. 1 退出
丁	4	1986. 1. 1 退出

解： $L = {}_2p_1 \cdot \mu_3 \cdot {}_4p_2 \cdot {}_2p_3 \cdot {}_2p_4$

$= e^{-4k}(3k)e^{-16k}e^{-8k}e^{-10k} = 3k \cdot e^{-38k}$

$$\ln L = \ln 3 + \ln k - 38k$$

$\frac{\partial \ln L}{\partial k} = 0$ 即 $\hat{k} = \frac{1}{38}$。■

（3）不完整数据、分组死亡、精确退出年龄。在精算实务中，经常遇到研究个人人身险投保人死亡率的情况。假设考虑期为 Z～Z+m 年（这里的“年”指保险年度），采用保险年龄以使所有被考察的投保人的年龄都为整数年龄，且进一步假设所有的

退出都发生在保险年度内，因而也是整数年龄。死亡年龄取为死亡那年对应的保险年龄。

设 n_x 为进入考察时年龄为 x 岁的人数，w_x 为在 x 岁退出的人数，θ_x 为在年龄区间 $(x, x+1]$ 内死亡的人数。观察记录的最低年龄为 a 岁，最高年龄为 h 岁。

显然，进入年龄区间 $(j, j+1]$ 的人数为：

$$A_j = \sum_{x=a}^{j}(n_x - w_x - \theta_{x-1}) = \sum_{x=a}^{j}(n_x - w_x - \theta_x) + \theta_j$$

则 $$A_j - \theta_j = \sum_{x=a}^{j}(n_x - w_x - \theta_x)$$

在这个年龄区间上，有 θ_j 个死亡，$A_j - \theta_j$ 个生存，这个年龄的似然因子为：

$$L_i = (q_j)^{\theta_j} \ (1 - q_j)^{A_j - \theta_j}$$

总似然函数为：

$$L = \prod_{j=a}^{h} L_j = \prod_{j=a}^{h}(q_j)^{\theta_j}(1 - q_j)^{A_j - \theta_j}$$

则 $$\ln L = \sum_{j=a}^{h}\theta_j \ln q_j - \sum_{j=a}^{h}(A_j - \theta_j)\int_j^{j+1}\mu_s ds$$

若 μ_s 的参数形式已知，则由 $\ln L$ 极大化此可获得参数估计。

11.3.3 分段参数修匀

前面讨论的参数形式都是单一的函数形式，但这样做很难得到满意的修匀结果。更一般的做法是，在 x 的不同区域上，用不同形式的含参函数拟合它。这种方法叫**分段函数修匀**，又叫**样条修匀**。

样条修匀的基本特征是：那些不同区域上的拟合函数有比较简单的形式，如二次或三次多项式。在结点处，满足光滑性要求，如连续、一次可微等。

1. 最小二乘三次样条。假设有初始估计值 u_x，$x \in [a, b]$。关于 t_x 的先验观点是修匀值 v_x 是（或近似可表示成）一个三次多项式。令

$$v_x = c_1 + c_2 x + c_3 x^2 + c_4 x^3$$

利用最小二乘法估计参数 c_1，c_2，c_3，c_4。可构造函数

$$SS = \sum_{x=a}^{b} w_x(u_x - c_1 - c_2 x - c_3 x^2 - c_4 x^3)^2 \tag{11.3.10}$$

令 $\frac{\partial SS}{\partial c_i} = 0(i = 1,2,3,4)$。可得矩阵方程

$$\mathbf{X'WXc} = \mathbf{X'Wu} \tag{11.3.11}$$

其中，

$$X=\begin{bmatrix}1 & a & a^2 & a^3\\ 1 & a+1 & (a+1)^2 & (a+1)^3\\ \vdots & \vdots & \vdots & \vdots\\ 1 & b & b^2 & b^3\end{bmatrix}_{(b-a-1)\times 4}$$

$$W=\begin{bmatrix}w_a & & & \\ & \ddots & & \\ & & w_{b-1} & \\ & & & w_b\end{bmatrix}_{(b-a-1)\times(b-a-1)}$$

$$c=(c_1,\ c_2,\ c_3,\ c_4)$$

求得 c ，完成修匀。

2. 两弧三次样条。关于 t_x 先验观点是：v_x 是 $[a,\ b]$ 上的分段三次多项式，在 k 处相连接，即

$$v_x=\begin{cases}p_0(x), & a\leqslant x\leqslant k\\ p_1(x), & k\leqslant x\leqslant b\end{cases}\tag{11.3.12}$$

这里，k 称为结点。

根据最小二乘法构造函数：

$$SS=\sum_{x=a}^{h}w_x[u_x-p_0(x)]^2+\sum_{x=h+1}^{b}w_x[u_x-p_1(x)]^2\tag{11.3.13}$$

这里 h 为不大于 k 的最大样本下标值 x。

为了得到光滑连接，要求在两个函数的连接处 k，有相同的函数值和一阶、二阶导数，即

$$\begin{cases}p_0(k)=p_1(k)\\ p_0'(k)=p_1'(k)\\ p_0''(k)=p_1''(k)\end{cases}\tag{11.3.14}$$

式（11.3.14）是确保在整个区间 $[a,\ b]$ 上的 t_x 二阶连续可微。由此可令

$$\begin{aligned}p_0(x)&=c_1+c_2x+c_3x^2+c_4x^3\\ p_1(x)&=c_1+c_2x+c_3x^2+c_4x^3+c_5(x-k)^3\end{aligned}\tag{11.3.15}$$

代入式（11.3.13）并令 $\dfrac{\partial SS}{\partial c_i}=0,(i=1,2,\cdots,5)$，可得关于 c 的线性方程组，并完成修匀。

一般的情形是存在 n 个结点 $k_1,\cdots,k_n$，v_x 由 n 个多项式组成，即

$$v_x=\begin{cases}p_0(x), & a\leqslant x\leqslant k_1\\ \vdots & \vdots\\ p_i(x), & k_i\leqslant x\leqslant k_{i+1}\\ \vdots & \vdots\\ p_n(x), & k_n\leqslant x\leqslant b\end{cases}$$

全部的光滑性（二阶连续可微）条件可表示成：

$$\begin{cases} p_{i-1}(k_i) = p_i(k_i) \\ p'_{i-1}(k_i) = p'_i(k_i), \qquad i=1,\cdots,n \\ p''_{i-1}(k_i) = p''_i(k_i) \end{cases}$$

由此可设

$$p_i(x) = p_0(x) + c_5\,(x-k_1)^3 + \cdots + c_{i+4}(x-k_i)^3, i=1,2,\cdots,n$$

在实际应用中，结点的选择非常重要。仔细观察 u_x 的图像，在形状有较大起伏的地方，应设置结点。结点个数越多，拟合程度越高，计算量也越大。

【例 11 – 6】　已知：

（1）$v_x = \begin{cases} p_0(x), & 10 \leqslant x \leqslant k \\ p_0(x) - 0.0064(x-k), & k \leqslant x \leqslant 13 \quad 11 < k < 12 \end{cases}$

（2）相关数据如表 11 – 3 所示。

表 11 – 3　　　　例 11 – 6 中的数据

x	w_x	u_x	$p_0(x)$
10	4	0.02	0.020
11	4	0.03	0.031
12	2	0.04	0.042
13	3	0.04	0.053

用最小二乘法样条修匀法来拟合观察值 u_x，并求 k。

解： $SS = \sum (v_x - u_x)^2 w_x$

$$= (0.02-0.02)^2 \times 4 + (0.031-0.03)^2 \times 4 + [0.042 - 0.0064(12-k) - 0.04]^2 \times 2 + [0.053 - 0.0064(13-k) - 0.04]^2 \times 3$$

令 $\frac{\partial SS}{\partial k} = 0$，得 $k = 11.3$。 ■

11.3.4　光滑连接修匀

样条修匀是用最小二乘法确定每一段（子区间）上的多项式函数。本小节仍然试图确定每一段（等距区间）上的多项式函数，但是，所用的方法却不是最小二乘法而是采用 Everett 公式，这种修匀方法叫做**光滑连接修匀**。

1. 表达式。光滑连接修匀方法的基本表达式如下：

$$v_{x+s} = F(s)u_{x+1} + F(t)u_x, \qquad 0 \leqslant s \leqslant 1,\ t = 1-s \qquad (11.3.16)$$

这个表达式又叫做 **Everett 公式**。

这里，$F(s) = A(s) + B(s)\delta^2 + C(s)\delta^4 + \cdots$，而 $A(s), B(s), C(s), \cdots$，是

关于 s 的函数。δ 是中心差分算子，其定义如下：

$$\begin{cases} \delta f(x) = f\left(x+\frac{1}{2}\right) - f\left(x-\frac{1}{2}\right) \\ \delta^2 f(x) = \delta[\delta f(x)] = f(x+1) - 2f(x) + f(x-1) \\ \vdots \\ \delta^n f(x) = f\left(x+\frac{n}{2}\right) - C_n^1 f\left(x+\frac{n}{2}-1\right) + C_n^2 f\left(x+\frac{n}{2}-2\right) + \cdots + (-1)^n f\left(x-\frac{n}{2}\right) \end{cases} \tag{11.3.17}$$

当 $F(s)$ 的表达式中的最后一项为 δ^{2m} 时，则在区间 [x, x+1] 内，v_{x+s} $(0 \leqslant s \leqslant 1)$ 的值除依赖于 s 的值外，还依赖于 2m+2 个 u_x 的值：

$$u_{x-m},\ u_{x-m+1},\ \cdots,\ u_{x+m+1}$$

因此，Everett 公式又叫 $(2m+2)$ 点插值公式。

当 $F(s) = s$ 时，式 (11.3.16) 变为：

$$v_{x+s} = su_{x+1} + tu_x$$

这便是大家熟悉的线性插值公式。

2. 光滑性。最基本的要求是：在结点处的修匀值相等，即

$$v_{x-1+s}\big|_{s=1} = v_{x+s}\big|_{s=0} \tag{11.3.18}$$

由 Everett 公式

$$\begin{aligned} v_{x-1+s}\big|_{s=1} &= A(1)u_x + B(1)\delta^2 u_x + C(1)\delta^4 u_x + \cdots \\ &\quad + A(0)u_{x-1} + B(0)\delta^2 u_{x-1} + C(0)\delta^4 u_{x-1} + \cdots \\ v_{x+s}\big|_{s=0} &= A(0)u_{x+1} + B(0)\delta^2 u_{x+1} + C(0)\delta^4 u_{x+1} + \cdots \\ &\quad + A(1)u_x + B(1)\delta^2 u_x + C(1)\delta^4 u_x + \cdots \end{aligned}$$

比较式 (11.3.18) 的两端，欲使之成为等式，当且仅当

$$A(0) = B(0) = C(0) = \cdots = 0 \tag{11.3.19}$$

若光滑性要求是在结点处的一阶导数相等，则

$$v'_{x-1+s}\big|_{s=1} = v'_{x+s}\big|_{s=0} \tag{11.3.20}$$

满足式 (11.3.18) 和 (11.3.20) 的插值公式称为**相切的**。

由于 $F'(s) = A'(s) + B'(s)\delta^2 + C'(s)\delta^4 + \cdots$

$$\frac{dF(t)}{ds} = -F'(t)$$

则比较式 (11.3.20) 两端，有

$$F'(1)u_x - F'(0)u_{x-1} = F'(0)u_{x+1} - F'(1)u_x \tag{11.3.21}$$

或　　$2F'(1)u_x = F'(0)(u_{x+1} + u_{x-1}) = F'(0)(2+\delta^2)u_x$

因此，　$2F'(1) = F'(0)(2+\delta^2)$　　(11.3.22)

若光滑性要求是在结点处二阶左、右导数相等，则

$$v''_{x-1+s}\big|_{s=1} = v''_{x+s}\big|_{s=0} \tag{11.3.23}$$

可推知：

$$F''(0)(u_{x+1}-u_{x-1})=0 \qquad (11.3.24)$$

则

$$F''(0)=0 \qquad (11.3.25)$$

$$A''(0)=B''(0)=C''(0)=0 \qquad (11.3.26)$$

即式（11.3.22）和（11.3.26）是插值公式密切的充分必要条件。

3. 再生性。若 Everett 公式具有再生性，则

$$v_x=u_x=v_{x+s}\big|_{s=0}=A(1)u_x+B(1)\delta^2u_x+C(1)\delta^4u_x+\cdots \qquad (11.3.27)$$

等价于

$$A(1)=1,B(1)=C(1)=\cdots=0 \qquad (11.3.28)$$

4. 精确度。Everett 公式的精确度指的是再生多项式函数的最高次数。

为了给出精确度的判别准则，先研究中心差分算子 δ 与向前差算子 Δ 的关系。由简单的代数运算，得恒等式

$$\delta^{2m}u_x=\Delta^{2m}u_{x-m} \qquad (11.3.29)$$

Δ 算子本身有如下性质：

$$\Delta^{2m}u_{y+1}=\Delta^{2m}u_y+\Delta^{2m+1}u_y \qquad (11.3.30)$$

从而推出 Everett 公式的向前差分表示：

$$\begin{aligned}v_{x+s}=&[A(s)+A(t)]u_x+A(s)\Delta u_x+[B(s)+B(t)]\Delta^2u_{x-1}\\&+B(s)\Delta^3u_{x-1}+[C(s)+C(t)]\Delta^4u_{x-2}+C(s)\Delta^5u_{x-2}+\cdots\end{aligned} \qquad (11.3.31)$$

由此可推出，Everett 公式的精度为 z 的充要条件是：表 11－4 所得到的前（$z+1$）个等式成立。

表 11－4

z	条　件
0	$A(s)+A(1-s)=1$
1	$A(s)=s$
2	$B(s)+B(1-s)=\frac{1}{2}s(s-1)$
3	$B(s)=\frac{1}{6}s(s^2-1)$
4	$C(s)+C(1-s)=\frac{1}{24}s(s^2-1)(s-2)$

5. 四点修匀公式。通常所说的四点修匀公式，其 $F(s)$ 的最后一项为 δ^2（$m=1$）。我们试图寻求满足如下条件的四点公式：

（1）精确度为 1；

（2）光滑性条件为结点处一阶导数相等，即是相切的。

此时，

$$v_{x+s}=F(s)u_{x+1}+F(t)u_x \qquad (11.3.32)$$

其中，$F(s)=A(s)+B(s)\delta^2$。由（1）知 $A(s)=s$，由（2）知 $B(0)=0$ 且 $2F'(1)=F'(0)(2+\delta^2)$，即

$$B(0)=0,B'(0)=0,B'(1)=\frac{1}{2} \qquad (11.3.33)$$

若 $B(1)=L$，且选择 B 为满足这些条件的最低次数多项式，则

$$B(s)=\left(3L-\frac{1}{2}\right)s^2+\left(\frac{1}{2}-2L\right)s^2 \qquad (11.3.34)$$

若取 $L=0$，则

$$B(s)=\frac{1}{2}s^2(s-1)$$

该公式称为 **Karup – King** 公式。

当 $L=\frac{1}{4}$时，$B(s)=\frac{1}{4}s^2$；

当 $L=\frac{1}{6}$时，$B(s)=\frac{1}{6}s^3$。

此公式是四点公式中的唯一密切公式，即在结点处 v_{x+s}的二阶左、右导数相等。

【例 11 –7】 用 Everett 四点修匀公式修匀 u_x 得到 v_x，已知

（1） 此公式是线性的；

（2） 此公式是密切的；

（3） $B(s)$是次数不超过 3 的多项式；

（4） $v_{x+\frac{1}{2}}=\sum\limits_{k=1}^{2}a_k u_{x+k}$。

求系数 a_2。

解： 由（1）知 $A(s)=s$，由（2）、（3）知 $B(s)=\frac{1}{6}s^3$。

$$v_{x+\frac{1}{2}}=F\left(\frac{1}{2}\right)u_{x+1}+F\left(\frac{1}{2}\right)u_x=\left[A\left(\frac{1}{2}\right)+B\left(\frac{1}{2}\right)\delta^2\right](u_{x+1}+u_x)$$

$$=\left(\frac{1}{2}+\frac{1}{48}\delta^2\right)(u_{x+1}+u_x)=\frac{1}{2}(u_{x+1}+u_x)+\frac{1}{48}(u_{x+2}-2u_{x+1}$$

$$+u_x+u_{x+1}-2u_x+u_{x-1})$$

把上式与（4）相比较易知，$a_2=\frac{1}{48}$。 ■

习 题

1. 什么是修匀？当修匀死亡率时，对修匀结果有些什么基本要求？

2. 一组初始估计及对应的修匀值如表 11 –5 所示。

表 11 –5

x	70	71	72	73	74	75	76	77	78
u_x	0.044	0.084	0.071	0.076	0.040	0.104	0.160	0.058	0.110
v_x	0.050	0.054	0.058	0.062	0.067	0.072	0.077	0.083	0.091

(1) 试计算光滑算子 $F_1 = \sum_x \Delta^3 v_x$ 的值。

(2) 试计算拟合算子 $F_2 = \sum_x w_x(u_x - v_x)$, $F_3 = \sum_x x w_x(u_x - v_x)$

这里 $w_x = n_x$ 的取值如表 11-6 所示。

表 11-6

x	70	71	72	73	74	75	76	77	78
w_x	135	143	140	144	149	154	150	139	145

(3) 若线性修正修匀值，即令 $v_x' = av_x + b$，使 $F_2 = F_3 = 0$，试确定 a，b 的值，并解释这种做法的直观含义。

3. 设有数据 u_x，$x = 20, 21, \cdots, 59$，若采用 $n = 6$ 的 M-W-A 方法修匀此数据组，则

(1) 计算 v_{42} 时，与 a_{-4} 相乘的 u_x 的下标值 x 为多少？

(2) 计算 v_{34} 时，与 u_{37} 相乘的系数 a_r 的下标值是多少？

(3) 有多少项用来计算某个特定的 v_x？

(4) 计算 v_x，v_{x+1} 和 v_{x+2} 时，将出现多少个不同的 u_x？

(5) 根据这些 u_x，能产生一些修匀值 v_x，确定这些 v_x 的下标 x 的范围。

4. U_1，U_2，U_3 是三个二项比例随机变量，假设它们相互独立，且有相同的方差 σ^2。设 t_x 是一个线性函数，分别从容量 n_1，n_2，n_3 的样本中，得到初始估计值 u_1，u_2，u_3。甲想采用重复实验且把容量扩大成 $3n_x$ 的方法产生 t_x 的较好的估计；乙想把 u_1，u_2，u_3 加权平均，即 $\sum_{i=1}^{3} \frac{n_i}{n_1 + n_2 + n_3} u_i$ 作为 t_x 的修正估计。试问谁的关于 t_x 的修正估计有较小的方差。

5. 若初始估计的暴露数如表 11-7 所示。

M-W-A 的具体表达式如下：

$$v_{44} = \sum_{r=-1}^{1} a_r u_{44-r}$$

它再生一次多项式，且在下列假设下最小化 $Var(v_{44})$ 以确定修匀值 v_{44}：

(1) 随机误差 $\{E_{x+r}\}$ 相互独立；

(2) $Var(E_{x+r}) = Var(E_x)\dfrac{n_x}{n_{x+r}}$，试求 a_0 及 v_{44} 的值。

表 11-7

x	43	44	45
n_x	125	150	150

6. 两种不同的 M-W-A 表达式如下：

$$V_x^I = \frac{1}{2}U_{x-1} + \frac{1}{2}U_{+1}$$

$$V_x^{II} = bU_{x-1} + aU_x + bU_{+1}$$

且已知：

(1) V_x'' 再生线性函数；

(2) $6Var(V_x') = Var(V_x'')$；

(3) 随机误差 E_{x+r} 相互独立，且方差相等；

(4) $a>0$。

试求 b 的值。

7. 考虑 $z=1$ 的 Whittaker 修匀：

(1) 假设对所有的 x，$w_x = w$ 为一常数，且 $h=w$，试比较 $w=10$ 时 v_x 的值与 $w=1$ 时 v_x 的值可得出什么结论？

(2) 假设对所有的 x，$w_x = w$ 为一常数，且 $h \neq w$，设 v_x 是在这种情形下的修匀结果。设对所有的 x 都有 $w_x = 1$ 且 $h^* = \dfrac{h}{w}$。此时的修匀结果为 v_x^*。试比较 v_x 和 v_x^*。

8. 在 Whittaker 修匀中，假设由 $\partial M/\partial v_n = 0$ 所得的后一个线性方程是 $-v_{n-3} + 3v_{n-2} - 3v_{n-1} + kv_n = u_n$，相应的权是 $w_n = 2$，求 k 的值。

9. 某个 Whittaker 修匀，可由下面方程组确定：$\mathbf{Cv} = \mathbf{wu}$。$\mathbf{u} = [1,\ 1,\ 2]'$；$h=3$；$z=1$；$w_x = 1$，$x=1,\ 2,\ 3$。试求 v_2。

10. 设 T_1，T_2，T_3 是相互独立的正态随机变量，其先验的均值和协方差分别是

$$\mathbf{m} = [2,\ 6,\ 6],\quad \mathbf{A} = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

观察值（初始估计值）为 $\mathbf{u} = [1,\ 14,\ 18]'$，条件协方差矩阵 $Var(\mathbf{U}\mid\mathbf{T})$ 为：

$$B = \begin{bmatrix} 1/5 & 0 & 0 \\ 0 & 3/4 & 0 \\ 0 & 0 & 2/3 \end{bmatrix}$$

试用 **K－J** 法求修匀向量 $\mathbf{v}$。

11. 对某 3 个观察值，用 **K－J** 法修匀，且已知：

(1) $Var(T_i) = i$，T_i 与 T_j 的相关系数为 $c_{ij} = \left(\dfrac{1}{2}\right)^{i-j}$，$i \neq j$；

(2) 矩阵 $\mathbf{B}$ 是对角阵，且对角元素就是 $\mathbf{A}$ 的对角元素；

(3) $\mathbf{v} - \mathbf{u} = [0.01,\ 0.02,\ 0.04]'$。

求 $m_1 - u_1$。

12. 用 Dirichlet 方法进行修匀，第 i 个时间区间的死亡率是先验死亡率 m_i 与估计死亡率的加权平均，先验死亡率之权数为 5/9。且 $m_i = 0.06$，$d = 200$，先验分布之参数分别为 $a_1, \cdots, a_n$，试求 a_1。

13. $v_{[x]}$ 是选择年龄为 x 的首年选择死亡率的修正值，v_x 是终极年龄为 x

的终极死亡率的修正值。假设这两列修匀值都是满意的。$\theta_{[x]+r}$表示选择年龄为 x，到达年龄为 $x+r$ 的观察死亡人数，$n_{[x]+r}$表示相应的暴露人数，假设它们是有效的。对于 $r=1$，$2,\cdots,k-1$，则关系式

$$v_{[x]+r}=a_r v_{[x+r]}+(1-a_r)v_{x+r},0<a_r<1$$

确定每个修正选择死亡率 $v_{[x]+r}$，这里 a_r 仅依赖于保障期限 r，而与 x 无关。这就是说，每个修匀选择死亡率是以下两个量的线性插值，它们是具有相同到达年龄的首年死亡率和终极死亡率，位于"向上对角线"的两端。对每个 r，a_r 由以下等式确定

$$\sum_x \theta_{[x]+r} = \sum_x n_{[x]+r}v_{[x]+r}$$

即保障期限同为 r，所有选择年龄的观察死亡人数的总和等于死亡人数的总和。

（1）选择期的长度是多少？

（2）当 r 增加时，在正选择模型中，你认为 a_r 将是上升还是下降？在负选择模型中呢？

（3）导出确定 a_r 的公式。

14. 对于表 11－8 的数据，（1）采用 Gompertz 形式修匀，确定 c 和 g。（2）将上题的 q_x 转换为 μ_x，应用最小二乘法重新确定 c。

表 11－8

x	65	66	67	68	69
q_x	0.02875	0.03099	0.03339	0.03598	0.03876

15. 对于表 11－9 的数据，采用 Weibull 形式修匀，其中 $a=0$，确定 n 和 k 的值。

表 11－9

x	30	31	32	33	34
μ_x	0.00065	0.00076	0.00089	0.00104	0.00121

16. 假设死亡力被认为是年龄的线性函数，$\mu_x=bx$。考虑由 n 个观察年龄同为 x 岁的人组成的样本，且考察到每个人均死亡为止。设第 i 个人的死亡年龄为 y_i，试求 b 的极大似然估计。

17. 对于 20 个确切年龄为 2 岁的人，假设死亡力具有如下形式：

$$\mu_x=\frac{1}{b+x},\ b\leqslant 0,\ x\geqslant 2$$

1 个死亡，19 个退出分别发生在确切年龄 3 岁，求 b 的极大似然估计。

18. 假设需要在区间［20，89］上拟合一个两次样条，设有 3 个接点：

$x=29.7$，62.5，80.6，在二阶连续可微光滑性条件下，写出修匀过程。

19. 若 Everett 公式为

$$v_{x+t}=A(s)u_{x+1}+B(s)\delta^2u_{x+1}+C(s)\delta^4u_{x+1}+\cdots$$
$$+A(1-s)u_x+B(1-s)\delta^2u_x+C(1-s)\delta^4u_{x+1}$$

试把中心差分算子换成向前差分算子。

20. 若 $u_{x-1}=4$，$u_x=7$，$u_{x+1}=15$，试求 $F(0)u_x$。

21. 考虑两点公式，其中 $A(s)=4s^3-3s^4$：

(1) 这个公式是相切的还是密切的；

(2) 这个公式是光滑的还是再生的。

第十二章 信度理论

学习目标

- ☐ 熟悉各种信度模型中信度估计的计算方法
- ☐ 掌握有限波动信度中的完全信度条件、部分信度和信度保费
- ☐ 掌握使用贝叶斯方法计算风险的后验分布和贝叶斯信度估计值
- ☐ 掌握应用 Bühlmann 模型、Bühlmann – Straub 模型及其推广，并理解它们与贝叶斯模型的关系
- ☐ 熟悉使用经验贝叶斯方法估计非参数、半参数和参数模式下的结构参数并计算信度估计值

§12.1 引 言

让我们简单回顾一下保费厘定的过程。保险人会首先根据险种的特点和所有被保险人的损失数据计算得出一个基础费率，然后根据被保险人风险特征的不同，选择费率因子对所有的被保险人进行分级，制定出各级别（风险子集）的相对费率，用来衡量此风险子集的平均风险水平。然而对于个体风险的厘定，由于上述费率系统无法完美展现和区分每个保单的所有风险，比如两份各费率因子水平完全一致的保单也很有可能面临不同的风险，这些差异来自于那些费率系统没有考虑或者难以考虑的因素。也就是说，仅仅依靠手册费率没有办法完美刻画被保险人的风险。

因此，在费率厘定中，精算师往往需要参考被保险人在过去一段时间内的损失数据来预测其未来的风险成本。这里又存在一个问题，由于经验损失数据来自经验期内发生的保险事故，因此这些数据本身就包含了很大的随机波动，仅仅采纳这些历史数据来估计将来的风险也是不准确的。现在需要解决的问题是，经验数据所反映的被保险人的风险水平与风险子集平均水平的差别中，有多少是由于随机波动所引起的？有多少是由于被保险人真的优于或劣于风险子集平均水平而引起的？如何分配两者的比重？换句话说，被保险人的自身索赔经验的可信度是多少？

信度理论为此提供了一个很好的工具。**信度理论是研究如何合理利用先验信息和个体索赔经验来进行估计、预测及制定后验保费。**后验保费估计值可以下面公式来表示：

$$后验保费估计值 = z \times 经验值 + (1 - z) \times 先验值$$

其中 z（$0\leqslant z\leqslant 1$）称为**信度因子**，后验保费估计值称为**信度估计**。只有正确地选择信度因子 z，才能保证调整后的保险费接近于真实的风险水平。除了保费外，信度理论还可以用来估计索赔数、总索赔额、损失率、级别相对数等值。

在本章中我们将详细介绍信度理论中几种常见模型。首先讨论有限波动信度模型。有限波动信度模型是20世纪早期发展的方法，也称为**古典信度模型**，此方法的一个目的是限制经验数据中的随机波动对估计准确性的影响。我们将在12.2节中给出完全可信条件，并讨论完全信度条件不满足时的处理方法，也就是部分信度。其次讨论贝叶斯信度模型。在12.3节中将介绍信度理论中贝叶斯方法的应用，并推导出基于经验数据对将来的预测，也就是贝叶斯信度估计。然后讨论一致最精确信度模型。有限波动模型强调估计结果的稳定性，而一致最精确信度模型则着重强调估计结果的精确性。后者实际上是对贝叶斯信度估计值的一种最佳线性逼近值。在12.4节中将介绍 Bühlmann 信度模型和 Bühlmann – Straub 信度模型及其推广。最后讨论经验贝叶斯信度估计。上述理论在实际应用时需要根据数据对未知分布的参数进行估计。根据分布形式是否已知，估计方法将分为非参数方法、半参数方法和参数方法，本书将分别在12.5.1，12.5.2和12.5.3中介绍。

§12.2　有限波动信度

假设 X 是我们要预测的随机变量，在这里，X 可以是损失次数、损失强度、总索赔额，也可以是保费、级别费率、损失率等。经验观察值为 X_j，$j=1,\cdots,n$，假设 $\{X_j,\ j=1,2,\cdots,n\}$ 互相独立。对于经验观察值的背景有两种理解：一种情况是过去 n 期内的值，比如过去 n 个月的月损失额，或者过去 n 年每年的损失次数，而 X_j 为过去第 j 期的观察值；另一种情况的经验数据是某 n 个同质保单的观察值，比如 n 个同质保单上一期的理赔额，其中 X_j 为第 j 个保单的观察值。

假设 $E(X_j)=\xi$，$1\leqslant j\leqslant n$。即认为被保险人在每个时期的损失期望或者同质保单组合中每个保单的损失期望都是相同的，保险人的目标是确定 ξ 的值，用以预测下一期的数据。同样也假设 $Var(X_j)=\sigma^2$，对所有 j 都是一致的。根据保险人先验知识，我们可以获得 X 均值的先验估计 M，如根据过去的理赔数据拟合分布得到，或者根据其他具有类似风险水平的被保险人的经验确定。如果我们要估计的 X 是纯保费，M 常称为**手册纯保费**。经验数据 X_j，$j=1,\cdots,n$ 的平均值可以用 $\bar{X}=n^{-1}(X_1+\cdots+X_n)$ 来描述。容易知道，$E(\bar{X})=\xi$，$Var(\bar{X})=\sigma^2/n$。有三种可能的方法预测 ξ。第一种是忽

略过去的经验数据，直接令 $\xi = M$；第二种方法是忽略 M，直接使用经验数据，即令 $\xi = \bar{X}$；第三种方法就是取 M 和 $\bar{X}$ 的加权值，即 $z\bar{X} + (1-z)M$，我们称 z 为信度因子。可以看出第一种表述和第二种表述分别是第三种表述下 z 为 0 和 z 为 1 的特殊形式。

一般说来，过去观测数据越多，提供的信息越充分，经验数据可信度就越高。此时，根据大数法则，可以直接用 $\bar{X}$ 估计 ξ，z 的值就越接近于1，保险人据此足以对将来获得正确的估计。特别地，当 $z=1$，称经验数据具有完全信度。相反，过去数据越少，提供的信息越不充分，经验数据的可信度就越低。这时 z 的值接近 0，保险人则只能基于先验信息估计。特别地，当 $z=0$ 时，称为经验数据没有信度。当 $0<z<1$ 时，称为经验数据具有部分信度。

12.2.1 完全信度

当 $\bar{X}$ 与 ξ 足够接近的时候，我们可以采用 $\bar{X}$ 来预测 ξ，即取 $z=1$，达到完全信度条件。那么怎样衡量 $\bar{X}$ 与 ξ 的接近程度呢？我们用 $\xi - r\xi \leqslant \bar{X} \leqslant \xi + r\xi$，即 $|\bar{X} - \xi| \leqslant r\xi$ 来表述。注意，这里由于 $X_j, j=1,\cdots,n$ 都是随机变量，所以 $\bar{X}$ 也是随机变量。因此我们设定在一定概率 p 下 $|\bar{X} - \xi| \leqslant r\xi$ 总是成立时，称 $\bar{X}$ 与 ξ 足够接近，这时对数据赋予完全可信性（r 接近 0，p 接近 1）。用公式表述为：

$$P(-r\xi \leqslant \bar{X} - \xi \leqslant r\xi) \geqslant p, \qquad r>0, 0<p<1 \tag{12.2.1}$$

上式可写为：

$$P\left(\frac{-r\xi}{\sqrt{\sigma^2/n}} \leqslant \frac{\bar{X} - \xi}{\sqrt{\sigma^2/n}} \leqslant \frac{r\xi}{\sqrt{\sigma^2/n}}\right) \geqslant p, \quad r>0, 0<p<1$$

定义

$$y_p = \inf_y \left\{ P\left(\left| \frac{\bar{X} - \xi}{\sigma/\sqrt{n}} \right| \leqslant y \right) \geqslant p \right\} \tag{12.2.2}$$

如果 $\bar{X}$ 有连续分布函数，则（12.2.2）式中的"$\geqslant$"号可以换成"$=$"号，这时 y_p 满足

$$P\left(\frac{|\bar{X} - \xi|}{\sigma/\sqrt{n}} \leqslant y_p\right) = p \tag{12.2.3}$$

由此，如果

$$\frac{r\xi}{\sqrt{\sigma^2/n}} \geqslant y_p \tag{12.2.4}$$

我们说观察值满足完全可信条件。记 $n_0 = y_p^2/r^2$，则式（12.2.4）可表述为：

$$(C1)\qquad n \geqslant \frac{y_p^2}{r^2}\frac{\sigma^2}{\xi^2} = n_0\left(\frac{\sigma}{\xi}\right)^2 \qquad (12.2.5)$$

上述公式说明当经验观察值的个数 n 不小于 $n_0(\sigma/\xi)^2$ 时，满足完全可信条件。

完全可信条件也可以用总期望值来描述，将式（12.2.5）改写为：

$$(C2)\qquad n\xi \geqslant n_0\frac{\sigma^2}{\xi} \qquad (12.2.6)$$

条件（C1）给出了满足完全可信条件的最小观察值个数，而条件（C2）给出了观察值的最小总期望值。如果当 X 代表年索赔次数时，则（C1）给出的是经验观察期的个数，而（C2）给出了所有观察期内发生索赔的期望总次数。如果 X 代表单位时间同质保单的纯保费，则（C1）给出了最小的同质保单数，（C2）给出了所观察的同质保单最小的期望总索赔额。

类似地，我们还可以得到完全可信的其他等价条件：

$$(1)\quad \frac{\sigma}{\xi} \leqslant \frac{r}{y_p}\sqrt{n} = \sqrt{\frac{n}{n_0}} \qquad (12.2.7)$$

上式说明如果变量 X 的变差系数 $CV = \sigma/\xi$ 不大于 $\sqrt{n/n_0}$，则经验数据是完全可信的。

$$(2)\quad Var(\bar{X}) = \frac{\sigma^2}{n} \leqslant \frac{\xi^2}{n_0} \qquad (12.2.8)$$

这个公式表明当 $\bar{X}$ 的波动性（方差）在一定范围内时，则可以认为经验数据是具有完全信度的。

当 $\bar{X}$ 的分布未知时，我们难以确定 y_p 的值。然而当 n 足够大时，由中心极限定理，

$$Z = \frac{\bar{X} - \xi}{\sigma/\sqrt{n}} \to N\ (0,\ 1) \qquad (12.2.9)$$

这时，y_p 为正态分布的 $\frac{1+p}{2}$ 分位点。通常我们使用 $p = 0.9$，$r = 0.05$。此时，

$$y_{0.9} = \Phi^{-1}\ (0.95) = 1.645$$

则　$n_0 = (y_{0.9}/r)^2 = 1\ 082.41$。完全可信条件为：

$$n \geqslant 1\ 082.41\ (\sigma^2/\xi^2)$$

【例 12－1】 设 N 表示某个体保单在单位时间内索赔次数，N 服从泊松分布，$E(N) = 5$，N_1，N_2，…，N_n 为其在过去相互独立的 n 个观察期内

发生的索赔次数，观察期 n 至少为多少时才满足完全信度条件？所有观察期内总共有多少索赔事件发生时，才能使索赔次数数据完全可信？（$p=0.9$，$r=0.05$）。

解：由于 N 服从泊松分布，可以得到 $E(N)=Var(N)$，我们已经计算得出 $n_0=1\ 082.41$。

由条件（C1）知，完全可信的条件为：

$$n \geqslant n_0 \times \frac{Var(N)}{E^2(N)} = 1\ 082.41 \times \frac{5}{25} = 216.48$$

由条件（C2）知，$N_1, N_2, \cdots, N_n$ 完全可信的条件为：

$$n\lambda = n_0 \times \frac{Var(N)}{E(N)} = 1\ 082.41 \times \frac{Var(N)}{E(N)} = 1\ 082.41$$

所以至少经历 217 个观察期，或至少有 1 083 个索赔事件发生，才能使索赔次数数据完全可信。读者可以注意到当 N 服从泊松分布时，我们不需要泊松参数 λ 也可以计算出完全信度条件（C2）。■

【例 12－2】 设 X 表示个体保单一次索赔的索赔额，服从均值为 5 的指数分布，假定 $r=0.05$，$p=0.9$，假设独立同质保单组合的索赔额经验数据为 $X_1, \cdots, X_n$，求索赔额的完全可信条件（C1）和（C2）。

解：n 表示观察到的索赔额的样本个数，$E(X)=5$，$Var(X)=25$，则由条件（C1）和（C2），有

$$n \geqslant n_0 \frac{\sigma^2}{\xi^2} = 1\ 082.41 \times \frac{25}{5^2} = 1\ 082.41$$

$$n\xi \geqslant 1\ 082.41 \times \frac{\sigma^2}{\xi} = 1\ 082.41 \times \frac{25}{5} = 5\ 412.05$$

因此，当观察到的总索赔次数 n 大于 1 082 时，索赔额经验数据完全可信，当期望总索赔额不小于 5 412.05 时，索赔额经验数据完全可信。

上面两个例子分别说明了个体保单的索赔次数和索赔额的完全可信条件，下面我们考虑个体保单总索赔额的完全可信条件。首先回忆一下第六章的复合分布的知识。设 S 表示单位时期内的某保单总索赔额，N 表示单位时期内索赔发生次数，Y_i 表示第 i 次索赔额，假设每次索赔额 Y_i 独立同分布，且都与索赔次数 N 独立。则总索赔额为：

$$S = Y_1 + Y_2 + \cdots + Y_N$$

由第六章，我们已经知道：

$$E(S) = E(Y)E(N)$$

$$Var(S) = E^2(Y)Var(N) + E(N)Var(Y)$$

设 $S_1, \cdots, S_n$ 表示过去 n 个时期内某个体保单总索赔额的经验数据，由条件（C1）和（C2）可以得到总索赔额经验数据的完全可信条件如下：

$$n \geqslant n_0 \frac{Var(S)}{E(S)^2}$$

$$= n_0 \frac{E(Y)^2 Var(N) + E(N) Var(Y)}{[E(N)E(Y)]^2} \tag{12.2.10}$$

式（12.2.10）给出了为保证过去总索赔额完全可信所需要最小的观察期。

$$E(N)E(Y) \geqslant n_0 \frac{Var(S)}{E(S)} = n_0 \frac{E(Y)^2 Var(N) + E(N)Var(Y)}{E(N)E(Y)} \tag{12.2.11}$$

式（12.2.11）给出了完全可信所需要的 n 个时期内的期望总索赔额。

另外，由式（12.2.10）有

$$nE(N) \geqslant n_0 \frac{E^2(Y)Var(N) + E(N)Var(Y)}{E^2(Y)E(N)}$$

其中 $nE(N)$ 表示 n 个观察期内的期望总索赔数。实际运用中，这个条件是用实际发生的索赔次数来表示。设 N_i 表示第 i 个时期发生的索赔数，用 $\sum_{i=1}^{n} N_i$ 来代替过去 n 个时期内发生的期望总索赔数$nE(N)$。下面通过一个复合泊松分布的例子阐述上述模型。

【例 12－3】 设 $X_1, X_2, \cdots, X_n, i=1,\cdots,n$ 为某保单组在过去 n 年内总索赔额，X_i 独立同分布，服从复合泊松分布，$X_i = Y_1^{(i)} + Y_2^{(i)} + \cdots + Y_{N_i}^{(i)}$。$N_i$ 为第 i 年的索赔次数，$N_i, i=1,\cdots,n$ 相互独立，服从参数为 500 的泊松分布，索赔强度为 Y，$Y_j^{(i)}$ 相互独立，表示第 i 期内第 j 次索赔额，服从均值为 5 的指数分布。$j=1,\cdots,N_i$。求（1）满足使总索赔额完全可信条件的最小观察年数，总索赔次数和总索赔额；（2）满足使索赔次数完全可信的最小观察年数和期望总索赔次数；（3）满足使索赔强度完全可信的最小总索赔次数和期望总索赔金额。

解：（1）记 $E(Y)=\theta_Y=5$，$Var(Y) = \sigma_Y^2 = 25$，则 $E(X_j) = \xi_X = \lambda\theta_Y = 2\ 500$，$Var(X_j) = \sigma_X^2 = \lambda(\theta_Y^2 + \sigma_Y^2) = 25\ 000$。由条件（C1）有：

$$n \geqslant n_0 \frac{\sigma_X^2}{\xi_X^2} = \frac{n_0}{\lambda}\left(1 + \frac{\sigma_Y^2}{\theta_Y^2}\right) = 4.3$$

即观察年数必须大于 5，才能保证总索赔额经验数据完全可信。由条件（C2）有

$$n\lambda \geqslant n_0 \frac{\sigma_X^2}{\xi_X^2}\lambda = n_0\left(1 + \frac{\sigma_Y^2}{\theta_Y^2}\right) = 2\ 164.8$$

即过去 n 年内发生总索赔数必须不小于 2 165，才能保证总索赔额经验数据完全可信。由式（12.2.11）

$$n\lambda\theta_Y \geqslant n_0\left(\theta_Y + \frac{\sigma_Y^2}{\theta_Y}\right) = 10\ 822.2$$

过去 n 年内发生总索赔额必须大于 10 824.1，才能保证总索赔额的数据完全可信。

（2）为计算使每年索赔次数达到完全可信条件的最小观察年数，由条件（C1）可得：

$$n \geqslant n_0 \times Var(N)/E^2(N) = n_0 \times \frac{500}{500^2} = 2.2$$

即需要观察至少 3 年。

由条件（C2）可得：

$$n\lambda \geqslant n_0 \times \frac{Var(N)}{E(N)} = 1\ 082.41$$

即在观察期内至少需要有 1 083 个索赔事件发生，才能够说索赔频率经验数据完全可信。

（3）为使每次索赔金额的经验数据达到完全可信条件，根据条件（C1）和（C2），可得：

$$n\lambda \geqslant n_0 \frac{\sigma_Y^2}{\theta_Y^2} = 1\ 082.41 \times \frac{25}{5^2} = 1\ 082.41$$

$$n\lambda\theta_Y \geqslant 1\ 082.4 \times \frac{\sigma_Y^2}{\theta_Y^2} = 1\ 082.41 \times \frac{25}{5} = 5\ 412.05$$

因此当观察到的期望总索赔次数不小于 1 083 时，索赔强度经验数据完全可信；当期望总索赔额不小于 5 412.65 时，索赔强度经验数据完全可信。 ■

【例 12－4】 已知某医疗保险中每个被保险人的年度医疗花费经验平均为 175，经验标准差为 140。现有一个团体医疗保险，在第一年内有 100 个人投保，在第二年有 110 个人投保。两年内人均获赔 150 元，问这个团体经验数据是否满足完全可信条件（$r = 0.05$ $p = 0.9$）？

解： 根据已知条件，年度医疗花费的期望和标准差的先验估计分别为 175 和 140。根据条件（C1），可得完全可信条件为：

$$n \geqslant n_0 \frac{\sigma^2}{\xi^2} = 1\ 028 \times \frac{140^2}{175^2} = 692.6$$

即至少需要 693 个被保险人的损失数据，才能认为经验数据是完全可信的。而经验观察人数为 $n = n_1 + n_2 = 100 + 110 = 210$ 人。由于 $210 < 693$，所以我们认为从观察人数来看，经验数据不完全可信。 ■

12.2.2 部分信度

当上例中的情况发生，经验数据不完全可信的时候，我们就要采用经验数据与先验知识的加权值作为保费的信度估计，即

$$P_c = z\bar{X} + (1 - z)M \tag{12.2.12}$$

其中 $0 < z < 1$。此时称经验数据是部分可信的。

z 的取值有许多方法，本章的主要任务是确定信度因子 z 的取值。这里给出一个比较简单的方法：有限波动信度法。在完全可信条件中，我们通

过控制 $\bar{X}$ 的方差使得 $\bar{X}$ 与 ξ 足够接近。但是从式（12.2.8）可以看出，虽然无法保证 $\bar{X}$ 的方差足够小，但是控制 P_c 的方差却是可行的。选择 z 使得 P_c 的方差等于完全可信条件时方差的上界，即

$$\frac{\xi^2}{n_0} = Var(P_c) = Var(z\bar{X} + (1-z)M) = z^2 Var(\bar{X}) = z^2\frac{\sigma^2}{n} \quad (12.2.13)$$

因此如果 z 比 1 小，则有 $z = (\xi/\sigma)\sqrt{n/n_0}$，于是可以统一写为：

$$z = \min\left(\frac{\xi}{\sigma}\sqrt{\frac{n}{n_0}},\ 1\right) = \min\left(\sqrt{\frac{n}{n_F}},\ 1\right) \quad (12.2.14)$$

其中 n_F 是完全信度条件要求的最小观察值个数 $n_F = n_0(\sigma/\xi)^2$。

式（12.2.14）中的第一等式的解释是，信度因子 z 是完全可信条件所要求的变异系数（$\sqrt{n/n_0}$）与实际变异系数的比值。第二个等式的解释是信度因子 z 是完全可信条件所要求最小观察值个数与实际样本数的比值的平方根。因此有限波动信度法也称**平方根法则**。

【例 12-5】 根据例 12-4，求信度因子和总信度费用。

解： 由上例可知，$\bar{X} = 150$，$M = 175$。满足完全可信条件的观察人数 $n_F = 692.6$，现在的观察总人数为 $n = 110 + 100 = 210$。根据式（12.2.14），信度因子 $z = \sqrt{\frac{n}{n_F}} = \sqrt{\frac{210}{692.6}} = 0.55$

于是信度估计值为 $z\bar{X} + (1-z)M = 0.55\times150 + 0.45\times175 = 161.25$。 ■

上例给出了索赔强度部分信度估计的计算方法，我们在通过一个例子考察复合泊松模型的部分信度估计。

【例 12-6】 假设总索赔额 X 服从复合泊松分布，索赔次数的参数为 λ，个体索赔额的均值为 θ_Y，方差为 σ_Y^2，设 $X_1,\cdots,X_n$ 是 n 个时期内的总索赔额观察值，求其信度因子。

解： 利用复合泊松分布的先验均值和方差公式计算得到：

$$\xi = E(X) = \lambda\theta_Y$$

$$\sigma^2 = Var(X) = \lambda(\theta_Y^2 + \sigma_Y^2)$$

代入式（12.2.14）得：

$$z = \min\left(\frac{\xi}{\sigma}\sqrt{\frac{n}{n_0}},\ 1\right) = \min\left(\sqrt{\frac{\lambda n/n_0}{1+\sigma_Y^2/\theta_Y^2}},\ 1\right) \quad (12.2.15)$$

■

【例 12-7】 假设总索赔额 X 服从复合泊松分布，其中索赔额变量服从指数分布。某投保团体过去 n 个时期总索赔额分别为 X_1，X_2，…，X_n，$\{X_j\}$ 是相互独立。假设索赔次数数据的可信度因子为 0.9，求总索赔额数据 X_1，X_2，…，X_n 的信度因子。

解：设该团体在过去 n 个时期的索赔次数为 N_1，N_2，…，N_n，由于索赔次数服从参数为 λ 的泊松分布，因此由泊松分布的信度公式知 $z_1=\sqrt{\frac{n}{n_F}}=\sqrt{\frac{n}{n_0/\lambda}}=\sqrt{\frac{\lambda n}{n_0}}=0.9$，即 $\lambda n/n_0=0.81$。因为索赔额变量服从指数分布，所以 $\sigma_Y=\theta_Y$。设 X_1，$X_2\cdots$，X_n 的信度因子为 z，则由式（12.2.15）知，

$$z=\min\left(\sqrt{\frac{\lambda n/n_0}{1+\sigma_Y^2/\theta_Y^2}},\ 1\right)=\min\left(\sqrt{\frac{0.81}{1+1}},\ 1\right)=0.6364$$ ■

值得注意的是，虽然用有限波动方法可以得到简明的解决方案，但在理论上却不能自圆其说。首先，这些 X_j 的分布没有潜在的理论模型支持，因此没有明确证明为什么信度保费形式（12.2.12）是恰当的，并且更优于 M。其次，即使对特定模型而言式（12.2.12）是恰当的，但却对如何选择 r 和 p 没有提出任何建议。最后，有限波动方法没有考虑 ξ 和 M 之间的差别。当运用式（12.2.12）时，本质上是认为 M 能够精确地代表在没有任何其他信息的条件下某被保险人的期望损失，然而，M 本身常常也是一个估计值，因此并不够可靠。所以，正确的信度问题的提法应该是“与 M 相比，$\bar{X}$ 的可靠程度增加了多少？”而不是“$\bar{X}$ 有多可靠？”。

本章下面将介绍贝叶斯建模方法。它是基于特定被保险人的索赔历史数据，并指出过去数据与预期费率的厘定相关。在一定条件，可以证明信度保费形式（12.2.12）是合理的，只是 z 的取值为另一种形式。

§12.3 贝叶斯信度

12.3.1 预测分布

正如我们在上节中所提到的，虽然经验数据 $\bar{X}$ 和手册保费 M 的加权形式直观上有助于提高估计的精度，然而并没有明确的理论证明这种形式是恰当的。1967 年，Bühlmann 提出建立在贝叶斯理论基础上的最大精度信度模型，又称 Bühlmann 信度，给出了信度的理论基础。为了更好地理解 Bühlmann 信度，我们先简单介绍一下贝叶斯理论在信度中的应用。

假设我们已知过去的 n 个经验数据 $\vec{X}=\{X_1,X_2,\cdots,X_n\}$ 的观察值 $\vec{x}=\{x_1,\ x_2,\cdots,x_n\}$，希望预测将来的数据 X_{n+1}。此外，我们还知道被保险人风险的先验信息 θ。虽然在厘定级别费率时，我们已经考虑了被保险人的若干风险因子，然而这些风险因子无法完美刻画被保险人的风险。我们用 θ 描述那些剩余的、未被观察到的、影响风险水平的因素，这样风险子集的某

一个被保险人的风险水平可以通过一个参数 θ 来描述。θ 的取值随被保险人的不同而不同，也就是说，通过不同取值 θ，我们可以区分不同被保险人的风险水平的差异。例如，在车险中，可以用不同的值表示车主的责任心、驾驶技巧等风险特征。我们通常假设 θ 的存在性，但是我们还进一步假设 θ 是不可观察的，因此我们永远不知道 θ 的真实值。

由于 θ 随被保险人变化，因此，在风险子集内存在一个关于 θ 的概率分布 $\pi(\theta)$。每个被保险人的具体风险参数 θ 的值都是未知的，我们视 θ 为随机变量 Θ 的观察值。如果 θ 是标量，则 Θ 的累积分布函数 $\Pi(\theta)=P(\Theta\leqslant\theta)$ 可以看做风险子集中风险水平小于或等于 θ 的被保险人的比例。在本节中，我们假定 $\pi(\theta)$ 是已知的，并称其为**先验分布或结构分布**。

不同的 θ 对应不同的风险，由经验数据 $\vec{X}$ 体现为不同的损失水平。X_j 关于 θ 的条件分布密度为 $f_{X_j|\Theta}(x_j|\theta),j=1,\cdots,n$，$\vec{X}$ 关于 θ 的条件 $P(\vec{X}=\vec{x}|\Theta=\theta)=f_{\vec{X}|\Theta}(\vec{x}|\theta)$。这里用向量形式表示经验数据。过去损失与将来损失关于过去数据的联合条件概率函数表示为 $f_{\vec{X},X_{n+1}|\Theta}(\vec{x},x_{n+1}|\theta)$。

已知先验分布 $\pi(\theta)$ 及条件分布 $f_{\vec{X}|\Theta}(\vec{x}|\theta)$，则 θ 和 $\vec{X}$ 的联合分布

$$f_{\vec{X},\Theta}(\vec{x},\theta)=\pi(\theta)\times f_{\vec{X}|\Theta}(\vec{x}|\theta) \tag{12.3.1}$$

进而，对 Θ 的所有可能取值积分得到 $\vec{X}$ 的边际密度（如果 Θ 的先验分布是离散的，则求和）：

$$f_{\vec{X}}(\vec{x})=\int_{\Theta}f_{\vec{X}|\Theta}(\vec{x}|\theta)\times\pi(\theta)d\theta$$

或

$$f_{\vec{X}}(\vec{x})=\sum_{\Theta}f_{\vec{X}|\Theta}(\vec{x}|\theta)\times\pi(\theta) \tag{12.3.2}$$

同样，我们也可以求得 $\vec{X}$ 与 X_{n+1} 的边缘密度函数：

$$f_{\vec{X},X_{n+1}}(\vec{x},x_{n+1})=\int_{\Theta}f_{\vec{X},X_{n+1}|\Theta}(\vec{x},x_{n+1}|\theta)\times\pi(\theta)d\theta \tag{12.3.3}$$

进而，我们可以求出 X_{n+1} 关于 $\vec{X}$ 的条件分布，称为预测分布：

$$f_{X_{n+1}|\vec{X}}(x_{n+1}|\vec{x})=\frac{f_{\vec{X},X_{n+1}}(\vec{x},x_{n+1})}{f_{\vec{X}}(\vec{x})} \tag{12.3.4}$$

注意，当 $\vec{X}$ 与 X_{n+1} 关于 θ 条件独立时，$\vec{X}$ 与 X_{n+1} 的边缘密度函数可表示为：

$$f_{\vec{X},X_{n+1}}(\vec{x},x_{n+1})=\int_{\Theta}f_{\vec{X}|\Theta}(\vec{x}|\theta)\times f_{X_{n+1}|\Theta}(x_{n+1}|\theta)\times\pi(\theta)d\theta \tag{12.3.5}$$

于是预测密度可以写为：

$$\begin{aligned}f_{X_{n+1}|\vec{X}}(x_{n+1}|\vec{x})&=\frac{f_{X_{n+1},\vec{X}}(x_{n+1},\vec{x})}{f_{\vec{X}}(\vec{x})}\\&=\int_{\Theta}f_{X_{n+1}|\Theta}(x_{n+1}|\theta)\times\pi_{\Theta|\vec{X}}(\theta|\vec{x})d\theta\end{aligned} \tag{12.3.6}$$

这里，

$$\pi_{\Theta|\vec{X}}(\theta|\vec{x})=\frac{f_{\Theta,\vec{X}}(\theta,\vec{x})}{f_{X_{n+1}}(x_{n+1})}=\frac{f_{\vec{X}|\Theta}(\vec{x}\mid\theta)\times\pi(\theta)}{\int_{\Theta}f_{\vec{X}|\Theta}(\vec{x}\mid\theta)\times\pi(\theta)d\theta}$$

是风险参数 Θ 的后验分布，而预测分布是后验分布与条件分布的乘积的积分或求和。这在直观上也是很容易理解的，因为式（12.3.6）就是以经验数据为条件的全概率公式。

需要注意是，上式成立的条件，被保险人的经验损失和预测损失之间需要关于风险参数 θ 条件独立，这在实际中很常见。如果条件独立不能满足，我们不能直接应用式（12.3.6），而需要根据情况从式（12.3.4）推导。经验数据不需要与预测数据无条件独立，事实上具有相同风险特征的保单的损失数据之间也往往不是无条件独立的。

【例 12－8】 假设某保险公司承保的业务中，有 1/3 的保单的月损失次数服从均值为 4 的泊松分布，有 2/3 的保单月损失次数服从均值为 2 的泊松分布，且同一张保单月的损失次数是独立的。对于一个随机抽取的保单，已知观察到上个月的损失次数为 1。求这个月此保单损失 1 次的概率。

解：根据题意，先验分布

$$\pi(\theta)=\begin{cases}\dfrac{1}{3}, & \theta=4\\[2ex] \dfrac{2}{3}, & \theta=2\end{cases}$$

泊松参数为均值，条件分布为 $f_{\vec{X}|\Theta}(k|\theta)=\dfrac{\theta^k}{k!}e^{-\theta}$，$k$ 为非负整数。于是根据全概率公式求得上个月损失次数为 1 的概率为：

$$f_{X_1}(1)=\sum_{\Theta}f_{X_1}(x_1\mid\theta)\times\pi_{\Theta}(\theta)=\frac{2^1}{1!}\times e^{-2}\times\frac{2}{3}+\frac{4^1}{1!}e^{-4}\times\frac{1}{3}=0.205$$

同样我们可以求出上个月和这个月损失次数都为 1 的概率：

$$f_{X_1,X_2}(1,1)=\sum_{\Theta}f_{X_1,X_2}(x_1,x_2\mid\theta)\times\pi_{\Theta}(\theta)$$

由于对于同张保单，上月的损失次数与这个月的损失次数是独立的，所以上式写为：

$$\sum_{\Theta}f_{X_1}(x_1\mid\theta)\times f_{X_2}(x_2\mid\theta)\times\pi_{\Theta}(\theta)=\left(\frac{2^1}{1!}\times e^{-2}\right)^2\times\frac{2}{3}+\left(\frac{4^1}{1!}e^{-4}\right)^2\times\frac{1}{3}$$
$$=0.051$$

于是可以求出条件分布：

$$P(X_2=1\mid X_1=1)=\frac{f_{X_1,X_2}(1,1)}{f_{X_1}(1)}=\frac{0.051}{0.205}=0.247$$

于是，对于上个月损失 1 次的保单，这个月也损失一次的概率为 0.247。 ■

注意，我们也可以通过后验分布求预测分布。经计算其后验分布为：

$$\pi_{\Theta|X_1}(\theta|1)=\begin{cases}0.88, & \theta=2\\ 0.12, & \theta=4\end{cases}$$

进而，

$$P(X_2=1|X_1=1)=\sum_{\Theta}f_{X_2|\Theta}(x_2|\theta)\times\pi_{\Theta|X_1}(\theta|x_1)$$

$$=\left(\frac{2^1}{1!}\times e^{-2}\right)\times 0.88+\left(\frac{4^1}{1!}e^{-4}\right)^2\times 0.12=0.247$$

【例 12－9】 已知某险种保单的被保险人的每次索赔额服从均值为1/Θ的指数分布，Θ 为随机变量，服从均值为 0.004，方差为 4×10^{-6} 的伽玛分布。已知同一保单过去三次的索赔数据分别为 100，950 和 450。求 Θ 的后验分布以及此保单下一次的索赔额的预测分布（已知同一保单的每次索赔之间独立）。

解： 设伽玛分布的参数为 α，β，则由伽玛分布的期望 $\alpha\beta$，方差为 $\alpha\beta^2$，可得参数 $\alpha=4$，$\beta=0.001$。从而先验分布为：

$$\pi(\theta)=\frac{(\theta/\beta)^{\alpha}e^{-\theta/\beta}}{\theta\Gamma(\alpha)}=\frac{\theta^3e^{-1\,000\theta}1\,000^4}{6},\qquad \theta>0$$

由题知，条件概率服从指数分布，为 $f_{\vec{X}|\Theta}(\vec{x}|\theta)=\theta e^{-x\theta}$。由后验分布的公式得：

$$\pi_{\Theta|\vec{X}}(\theta|100,950,450)=\frac{f_{\vec{X},\Theta}(\vec{x},\theta)}{f_{\vec{X}}(\vec{x})}$$

$$=\frac{\theta^3e^{-(100+950+450)\theta}\times\theta^3e^{-1\,000\theta}1\,000^4/6}{f_{\vec{X}}(\vec{x})}$$

注意到分母为经验损失的边缘分布，是一个与参数 θ 无关的值，将其与分子中常数统一用常数 C 表示，则可以得到：

$$\pi_{\Theta|\vec{X}}(\theta|100,950,450)=C\times e^{-2\,500\theta}\theta^6\propto e^{-2\,500\theta}\theta^6$$

由于上式为密度函数，所以积分为 1，可求出上式的常数 C。又从 $\pi_{\Theta|\vec{X}}(\theta|100,950,450)$ 的形式可以看出，其对应的正好是参数为 7 和 0.0004 的伽玛分布，因此不需要积分，直接可以获得风险参数的后验分布为：

$$\pi_{\Theta|\vec{X}}(\theta|100,950,450)=\frac{2\,500^7\theta^6e^{-2\,500\theta}}{\Gamma(7)}$$

进而，根据式（12.3.6），可以求得预测分布：

$$f_{X_4|\vec{X}}(x|100,950,450)=\int_{\Theta}f_{X_4|\Theta}(x|\theta)\times\pi_{\Theta|\vec{X}}(\theta|100,950,450)d\theta$$

$$=\int_0^{\infty}\theta e^{-\theta x}\frac{\theta^6e^{-2\,500\theta}2\,500^7}{\Gamma(7)}d\theta$$

$$=\frac{2\,500^7\times 7}{(2\,500+x)^8}$$

服从参数为 7 和 2 500 的帕累托分布。

定义 12－1 如果参数 θ 的后验分布与先验分布具有相同的形式，只是参数不同，则称此先验分布为给定模型的共轭先验分布。

【例 12－10】 设 θ 的先验分布 $\pi(\theta)$ 为贝塔分布 $\beta(a,b)$，**X** 关于 **θ** 的条件分布为二项分布 $B(m,\theta)$，证明 $\beta(a,b)$ 是共轭先验分布。

解： 由于 $f_{X|\Theta}(x\mid\theta)=C_m^x\theta^x(1-\theta)^{m-x}$，$x=0,1,2,\cdots,m$

$$\pi(\theta)\propto\theta^{a-1}(1-\theta)^{b-1}$$

所以，

$$f(\theta,x)\propto\theta^{a-1}(1-\theta)^{b-1}\theta^x(1-\theta)^{m-x}=\theta^{a+x-1}(1-\theta)^{m+b-x-1}$$

由于 $f(x)$ 不受 θ 影响，因此，

$$f(\theta\mid x)=\frac{f(\theta,x)}{f(x)}\propto\theta^{a+x-1}(1-\theta)^{m+b-x-1}\sim\beta(a+x,m+b-x)$$

所以条件分布为二项分布时，$\beta(a,b)$ 是共轭先验分布。

关于共轭先验分布，可以证明下面的结论：

定理 12－1 设 $\{X_j,j=1,\cdots,n\}$ 关于 Θ 条件独立，分布密度函数具有如下的形式：

$$f_{X_j|\Theta}(x_j\mid\theta)=\frac{p(x_j)e^{-\theta x_j}}{q(\theta)}$$

且当参数 θ 取有限实值时，其先验分布满足如下形式：

$$\pi(\theta)=\frac{[q(\theta)]^{-k}e^{-\theta\mu k}}{c(\mu,k)}$$

其中 μ，k 为分布的参数。则参数 θ 后验分布与先验分布共轭。

表 12－1 是一些常见的共轭分布，请读者自己验证。

表 12－1 常见共轭分布

条件分布	共轭先验分布	后验分布
泊松分布 $P(\Theta)$	伽玛分布 (α,λ)	伽玛 $(\alpha+n\bar{x},n+\lambda)$
二项分布 $B(n,\Theta)$（随机参数为 Θ）	贝塔分布 $\beta(a,b)$	贝塔 $(a+x,n+b-x)$
均值为 Θ 的指数分布	伽玛分布 (α,λ)	伽玛 $(\alpha+n\bar{x},n+\lambda)$
均值为 $1/\Theta$ 的指数分布	逆伽玛分布 $I\Gamma(\alpha,\lambda)$	逆伽玛 $(\alpha+n\bar{x},n+\lambda)$

12.3.2 贝叶斯信度估计

现在回到前面的问题，已知过去的 n 个经验数据 $\vec{X}=\{X_1,X_2,\cdots,X_n\}$ 的观察值 $\vec{x}=\{x_1,x_2,\cdots,x_n\}$，预测下一期的损失数据 X_{n+1}。假设当风险参数 $\Theta=\theta$ 给定时的 X_{n+1} 的条件分布为 $f_{X_{n+1}|\Theta}(x_{n+1}\mid\theta)$，则最理想的情形是 θ 已知，下期的期望损失为条件期望 $\mu_{n+1}(\theta)=E(X_{n+1}\mid\Theta=\theta)$，这样得到的预

测值称为**个体估计值**。然而对每个特定的被保险人，θ 是不可观察的。若我们对 θ 和经验数据一无所知，则认为下期的期望损失为无条件期望，即 $E(X_{n+1})=E[E(X_{n+1}\mid\Theta)]$，这样得到的预测值称为**集体估计值**。一般情况下，我们对 θ 是有一定先验认识的。假设我们已知 θ 的先验分布 $\pi(\theta)$，并假设该被保险人的过去损失经验独立，即给定 $\Theta=\theta$ 时，X_1，$X_2,\cdots,X_n$ 是独立的，则由前面知，我们可以利用过去的经验数据对该被保险人明年的损失 X_{n+1} 进行估计，即使用预测分布的期望值

$$E(X_{n+1}\mid\vec{X}=\vec{x})=\int_0^{\infty}x_{n+1}\times f_{X_{n+1}\mid\vec{X}}(x_{n+1}\mid\vec{x})\,dx_{n+1}$$

这样得到的估计值，称为**贝叶斯信度估计值**。

进一步，由式（12.3.6），贝叶斯信度估计值可写做：

$$E(X_{n+1}\mid\vec{X}=\vec{x})=\int_{X_{n+1}}x_{n+1}\times f_{X_{n+1}\mid\vec{X}}(x_{n+1}\mid\vec{x})\,dx_{n+1}$$

$$=\frac{\int_{X_{n+1}}x_{n+1}\times f_{\vec{X},X_{n+1}}(\vec{x},x_{n+1})\,dx_{n+1}}{f_{\vec{X}}(\vec{x})}$$

也可以用预测分布写为：

$$E(X_{n+1}\mid\vec{X}=\vec{x})=\int_{X_{n+1}}x_{n+1}\times f_{X_{n+1}\mid\vec{X}}(x_{n+1}\mid\vec{x})\,dx_{n+1}$$

$$=\int_{X_{n+1}}x_{n+1}\times\int_{\Theta}f_{X_{n+1}\mid\Theta}(x_{n+1}\mid\theta)\times\pi_{\Theta\mid\vec{X}}(\theta\mid\vec{x})\,d\theta dx_{n+1}$$

$$=\int_{\Theta}\int_{X_{n+1}}x_{n+1}\times f_{X_{n+1}\mid\Theta}(x_{n+1}\mid\theta)\,dx_{n+1}\times\pi_{\Theta\mid\vec{X}}(\theta\mid\vec{x})\,d\theta$$

$$=\int_{\Theta}E(X_{n+1}\mid\theta)\times\pi_{\Theta\mid\vec{X}}(\theta\mid\vec{x})\,d\theta \tag{12.3.7}$$

当 Θ 服从离散分布时，

$$E(X_{n+1}\mid\vec{X}=\vec{x})=\sum_{\Theta}E(X_{n+1}\mid\theta)\times\pi_{\Theta\mid\vec{X}}(\theta\mid\vec{x})$$

需要注意的是，式（12.3.7）与式（12.3.5）建立在同样的假设上：对于相同经验损失数据和预测损失关于风险参数的条件分布是独立的。

【例 12－11】 承例 12－8，求此被保险人这个月损失的预测分布和贝叶斯信度估计值。

解：我们已经求得后验分布：

$$\pi_{\Theta\mid X_1}(\theta\mid 1)=\begin{cases}0.88, & \theta=2\\ 0.12, & \theta=4\end{cases}$$

根据式（12.3.6）和（12.3.7），可以得到预测分布和贝叶斯信度估计值：

$$f_{X_2\mid X_1}(x\mid 1)=\sum_{\Theta}f_{X_2\mid\Theta}(x\mid\theta)\times\pi_{\Theta\mid X_1}(\theta\mid 1)$$

$$=0.88\times\frac{2^x}{x!}e^{-2}+0.12\times\frac{4^x}{x!}e^{-4}$$

$$E(X_2 \mid X_1 = 1) = \sum_{\Theta} E(X_{n+1} \mid \theta) \times \pi_{\Theta \mid X_1}(\theta \mid 1) = 2 \times 0.88 + 4 \times 0.12 = 2.24$$

【例 12－12】 承例 12－9，已知单笔索赔额服从均值为 $1/\Theta$ 的指数分布，所有被保险人的 Θ 服从均值为 0.004、方差为 4×10^{-6}的伽玛分布。已知同一保单过去三次的索赔数据分别为 100，950 和 450。求此保单下一次索赔额的贝叶斯信度估计值（已知同一保单的每次损失之间独立）。

解： 我们已经求得预测分布 $f_{X_4 \mid \vec{X}}(x \mid 100,\ 950,\ 450) = \dfrac{2\,500^7 \times 7}{(2\,500 + x)^8}$，服从参数为 7 和 2 500 的帕累托分布。根据帕累托分布取期望，得到贝叶斯信度估计值：

$$E[X_4 \mid \vec{X} = (100,\ 950,\ 450)] = \frac{2\,500}{7-1} = 416.67$$ ■

注意，若去掉风险参数 θ 的随机性，直接对条件期望 $E(X_{n+1} \mid \Theta)$ 再取期望，得到无条件期望为 $E(X_{n+1}) = E_{\Theta}[E(X_{n+1} \mid \Theta)] = E\left(\dfrac{1}{\Theta}\right) = 333.3$。另外，直接求经验数据的平均值，得到 $\bar{X} = 500$。比较贝叶斯信度估计值、集体估计值和个体经验均值，可以发现贝叶斯信度估计介于两者之间。

【例 12－13】 有某类团险保单组，在第 j 年有 m_j 个被保险人。假设每位被保险人的年索赔次数相互独立，且服从参数为 θ 的泊松分布。风险参数 Θ 服从伽玛分布：

$$\pi(\theta) = \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\Gamma(\alpha)\ \beta^{\alpha}},\quad \theta > 0$$

每人每次索赔额都为常数，每年以 $100r\%$ 的通货膨胀率递增。若第 0 年的每次索赔额为 c 元，则在第 t 年的每次索赔额为 $(1+r)^t c$。设各年每张保单的人均纯保费等于该年人均总索赔额期望值。已知保单组在第 $n+1$ 年该团体有 m_{n+1} 个被保险人，求人均纯保费的贝叶斯信度估计值。

解： 在第 j 年的保单组的总索赔次数为 N_j，服从参数为 $m_j\theta$ 的泊松分布，即

$$P(N_j = n \mid \Theta = \theta) = \frac{(m_j\theta)^n e^{-m_j\theta}}{n!},\qquad n = 0,\ 1,\ 2,\ \cdots$$

设 X_j 表示该投保团体在第 j 年的人均年总索赔额，

$$X_j = c\ (1+r)^j N_j / m_j,\quad j = 1, 2, \cdots, n$$

因此，X_j 的条件分布为：

$$f_{X_j \mid \Theta}(x_j \mid \theta) = P(N_j = c^{-1}(1+r)^{-j} m_j x_j \mid \Theta = \theta)$$

$$= \frac{[m_j\theta]^{c^{-1}(1+r)^{-j} m_j x_j} e^{-m_j\theta}}{[c^{-1}(1+r)^{-j} m_j x_j]!}$$

则 $(X_1, X_2, \cdots, X_n, \Theta)$ 的联合分布密度为：

$$f_{\vec{X},\Theta}(\vec{x}\mid\theta) = (\prod_{j=1}^{n} f_{X_j\mid\Theta}(x_j\mid\theta))\pi(\theta)$$

Θ 的后验分布为：

$$\pi_{\Theta\mid\vec{X}}(\theta\mid\vec{x}) = \frac{f_{\vec{X},\Theta}(\vec{x}\mid\theta)}{f_{\vec{X}}(\vec{x})}$$

$$= \frac{(\prod_{j=1}^{n}[m_j\theta]^{c^{-1}(1+r)^{-j}m_jx_j}e^{-m_j\theta})\theta^{\alpha-1}e^{-\theta/\beta}}{[c^{-1}(1+r)^{-j}m_jx_j]!\times\Gamma(\alpha)\beta^{\alpha}\times f_{\vec{X}(\vec{x})}}$$

$$= A\theta^{\alpha+c^{-1}\sum_{j=1}^{n}(1+r)^{-j}m_jx_j-1}e^{-\theta(\frac{1}{\beta}+\sum_{j=1}^{n}m_j)}$$

从后验分布的形式可以看出，$\pi_{\Theta\mid\vec{X}}(\theta\mid\vec{x})$是一个参数为

$$\alpha_* = \alpha + c^{-1}\sum_{j=1}^{n}(1+r)^{-j}m_jx_j \quad \beta_* = \left(\frac{1}{\beta}+\sum_{j=1}^{n}m_j\right)^{-1}$$

的伽玛分布。另外，X_j 的条件期望为：

$$E(X_j\mid\Theta=\theta) = E\left(\frac{c(1+r)^jN_j}{m_j}\mid\Theta=\theta\right) = c(1+r)^j\theta$$

因此，$n+1$ 年的人均风险保费为 $\mu_{n+1}(\theta)=c(1+r)^{n+1}\theta$，人均纯保费为：

$$\mu_{n+1} = E(\mu_{n+1}(\Theta)) = c(1+r)^{n+1}\alpha\beta$$

由式（12.3.7）计算人均贝叶斯信度估计值为：

$$E(X_{n+1}\mid\vec{X}=\vec{x}) = \int_0^{\infty}\mu_{n+1}(\theta)\pi_{\Theta\mid\vec{X}}(\theta\mid\vec{x})d\theta$$

$$= E(\mu_{n+1}(\theta)\mid\vec{X}=\vec{x})$$

$$= E(c(1+r)^{n+1}\Theta\mid\vec{X}=\vec{x})$$

$$= c(1+r)^{n+1}\alpha_*\beta_*$$

事实上，若令 $m=\sum_{j=1}^{n}m_j$ 为 n 年内该团体的投保总人数，则经过简单的代数运算可以得到：

$$E(X_{n+1}\mid\vec{X}=\vec{x}) = z\bar{x}+(1-z)\mu_{n+1}$$

其中，$z=m/(m+\beta^{-1})$，$\bar{x}=m^{-1}\sum_{j=1}^{n}(1+r)^{n+1-j}m_jx_j$。若将 $\bar{x}$ 视为该团体在 n 年内的经验保费，则贝叶斯信度估计值可以看做是该团体经验人均保费与人均纯保费的加权平均。 ■

贝叶斯信度估计的做法是用经验数据的信息代替风险参数，求$E(X_{n+1}\mid\vec{X}=\vec{x})$。从直观上，我们可以想象贝叶斯信度估计值优于条件期望的期望，因为它更加充分地利用了已知的信息。事实上，贝叶斯信度估计值是所有利用经验数据估计 X_{n+1}中误差最小的估计量，见下面的定理。

定理 12－2 在所有利用 $\vec{X}$ 来估计 X_{n+1}的估计量中，贝叶斯信度估计值

的均方误差最小，即 $E\{[X_{n+1}-E(X_{n+1}|\vec{X})]^2\}$ 小于任何其他估计量的均方误差。

证明：记 X_{n+1} 的任一估计量为 $g(\vec{X})$，则其均方误差为：

$$Q_g=E\{[X_{n+1}-g(\vec{X})]^2\}$$

贝叶斯信度估计值的均方误差为：

$$E\{[X_{n+1}-E(X_{n+1}|\vec{X})]^2\}=E\{X_{n+1}^2-2X_{n+1}E(X_{n+1}|\vec{X})+E^2(X_{n+1}|\vec{X})\}$$

$$=E(X_{n+1}^2)-2E[X_{n+1}E(X_{n+1}|\vec{X})]+E[E^2(X_{n+1}|\vec{X})] \qquad (12.3.8)$$

由期望的迭代法则，可将第二项

$$E[X_{n+1}E(X_{n+1}|\vec{X})]=E\{E[X_{n+1}E(X_{n+1}|\vec{X})]|\vec{X}\}$$

$$=E\{E(X_{n+1}|\vec{X})E(X_{n+1}|\vec{X})] \qquad (12.3.9)$$

注意最后一个等式成立是因为 $E(X_{n+1}|\vec{X})$ 可看做是随机向量 $\vec{X}$ 的函数，因此当对 $E(X_{n+1}|\vec{X})$ 关于 $\vec{X}$ 量条件期望时，由条件期望的性质①可得。

由式（12.3.8）和式（12.3.9）有

$$E\{[X_{n+1}-E(X_{n+1}|\vec{X})]^2\}=E(X_{n+1}^2)-E[E^2(X_{n+1}|\vec{X})]$$

$$=E(X_{n+1}^2)-2E[g(\vec{X})X_{n+1}]+2E[g(\vec{X})X_{n+1}]+E[g^2(\vec{X})]-E[g^2\ (\vec{X})]-E[E^2(X_{n+1}|\vec{X})]$$

$$=E\{[X_{n+1}-g(X)]^2\}-E[g^2(\vec{X})]+2E[g(\vec{X})X_{n+1}]-E[E^2(X_{n+1}|\vec{X})]$$

$$=E\{[X_{n+1}-g(\vec{X})]^2|-E\{[g(\vec{X})\ -E(X_{n+1}|\vec{X})]^2\}$$

$$=Q_g-E\{[g(\vec{X})-E(X_{n+1}|\vec{X})]^2\}$$

所以，$E\{[X_{n+1}-E(X_{n+1}|\vec{X})]^2\}\ \leqslant Q_g$ ■

虽然贝叶斯信度估计值相对而言比较精确，但是它的计算方法比较困难，计算过程中需要多次用到积分。虽然前两个例子都可以很快的计算出结果，但是实际应用中会遇到更加复杂的分布，有时还要用到数值积分的方法。为了得到较为简便的结果，Bühlmann（1967）② 提出用 $\vec{X}$ 的线性函数 $\tilde{X}_{n+1}=\alpha_0+\sum_{j=1}^{n}\alpha_jX_j$ 来逼近 $\mu_{n+1}(\theta)$，其中 α_0，α_1，…，α_n 为要确定的参数。选择系数的原则是使估计的均方误差最小化，即最小化

$$Q=E[\mu_{n+1}(\Theta)-(\alpha_0+\sum_{j=1}^{n}\alpha_jX_j)]^2 \qquad (12.3.10)$$

注意到这里的 Q 中的随机变量除了 $\vec{X}$ 还包含 Θ，因此式（12.3.10）

① 这里用到了条件期望的性质：设 X 和 Y 是任意两个可积随机变量，$g(\cdot)$ 是一个函数，使得 $E[g(Y)]$，$E[Xg(Y)]$ 存在，则 $E(Xg(Y)|Y)=g(Y)E(X|Y)$。关于这个性质的证明可参考严士健、刘秀芳：《概率与测度》，北京师范大学出版社 2003 年版，第 271 页。

② Bühlmann, H. (1967), "Experience Rating and Credibility," ASTIN Bulletin, 4, 199－207.

的期望是针对所有的 X_1，$X_2,\cdots,X_n$ 和 Θ。

为最小化式（12.3.10），我们可以通过对所有的待定参数 α_0，$\alpha_1,\cdots$，α_n 求偏导，令其等于0，得

$$0=\frac{\partial Q}{\partial\hat{\alpha}_0}=-2E[\mu_{n+1}(\Theta)]+2E[(\hat{\alpha}_0+\sum_{j=1}^{n}\hat{\alpha}_jX_j)] \qquad (12.3.11)$$

$$0=\frac{\partial Q}{\partial\hat{\alpha}_j}=-2E\{X_i[\mu_{n+1}(\Theta)-\hat{\alpha}_0-\sum_{j=1}^{n}\hat{\alpha}_jX_j]\},\ i=1,2,\cdots,n \qquad (12.3.12)$$

在上面的公式中，$E[\mu_{n+1}(\Theta)]=E[E[X_{n+1}\mid\Theta]]=E(X_{n+1})$

根据式（12.3.11），可以得到：

$$E(X_{n+1})=\hat{\alpha}_0+\sum_{j=1}^{n}\hat{\alpha}_jE(X_j) \qquad (12.3.13)$$

式（12.3.13）说明 $\tilde{X}_{n+1}$ 是 X_{n+1} 的无偏估计，通常称为无偏方程。式（12.3.12）中的

$$E[X_i\mu_{n+1}(\Theta)]=E\{E[X_i\mu_{n+1}(\Theta)\mid\Theta]\}=E[E(X_i\mid\Theta)E(X_{n+1}\mid\Theta)]$$

在每次损失条件分布独立的情况下，

$$E[X_i\mu_{n+1}(\Theta)]=E[E(X_iX_{n+1}\mid\Theta)]=E(X_iX_{n+1})$$

所以式（12.3.12）可以化简为：

$$E(X_iX_{n+1})=\hat{\alpha}_0E(X_i)+\sum_{j=1}^{n}\hat{\alpha}_jE(X_iX_j) \qquad (12.3.14)$$

把式（12.3.13）两边乘以 $E(X_i)$，然后与式（12.3.14）相减得：

$$Cov(X_i,X_{n+1})=\sum_{j=1}^{n}\hat{\alpha}_jCov(X_iX_j),\qquad i=1,2,\cdots,n \qquad (12.3.15)$$

上式组成正交方程组，可以写为：

$$Cov(\vec{X},X_{n+1})=\vec{\hat{\alpha}}_j\Sigma$$

其中 Σ 为协方差阵。Σ 非奇异时，$\vec{\hat{\alpha}}_j=Cov(\vec{X},X_{n+1})\Sigma^{-1}$。

下面两条定理将说明，这种 Bühlmann 线性估计不但是对 $E(X_{n+1}\mid\Theta)$ 的线性估计中均方误差最小估计，也是 X_{n+1} 和贝叶斯信度估计值的最小误差线性估计。

定理 12－3　Bühlmann 线性估计在所有用 $\vec{X}$ 对 X_{n+1} 的线性估计中均方误差最小。

证明：$\vec{X}$ 对 X_{n+1} 的任一线性估计为 $\alpha_0+\sum_{j=1}^{n}\alpha_jX_j$，其均方误差 $Q_1=E[X_{n+1}-(\alpha_0+\sum_{j=1}^{n}\alpha_jX_j)]^2$。对 Q_1 求偏导，得

$$\frac{\partial Q_1}{\partial\alpha_0}=-2E[X_{n+1}]+2E[(\alpha_0+\sum_{j=1}^{n}\alpha_jX_j)] \qquad (12.3.16)$$

$$\frac{\partial Q_1}{\partial\alpha_i}=2E\{X_i[X_{n+1}-\alpha_0-\sum_{j=1}^{n}\alpha_jX_j]\}$$

$$= 2E(X_iX_{n+1}) - 2[\alpha_0 E(X_i) + \sum_{j=1}^{n} \alpha_j E(X_iX_j)], \ i = 1,2,\cdots,n \quad (12.3.17)$$

显然二阶偏导数矩阵是正定的。从式（12.3.16）和（12.3.17）可以看出，当 $\alpha_i = \hat{\alpha}_i$，$i = 0,\cdots,n$ 时，$\frac{\partial Q_1}{\partial \alpha_i}$等于0。因此 Bühlmann 线性估计在所有用 $\vec{X}$ 对 X_{n+1}的线性估计中均方误差最小。

定理 12－4 Bühlmann 线性估计在所有用 $\vec{X}$ 对贝叶斯信度估计值的线性估计中均方误差最小的。

证明： 证明过程与定理 12－3 类似，请读者自行证明。

综上所述，上面得到的 Bühlmann 线性估计很好的性质，它对 $E(X_{n+1}|\Theta)$、X_{n+1} 和贝叶斯信度估计值的最好的线性估计，下节要介绍的 Bühlmann 信度模型及其扩展都是上述 Bühlmann 线性估计的特殊形式。

§12.4 最大精度信度模型

在上节中我们推导出了 Bühlmann 线性估计是对 $E(X_{n+1}|\Theta)$、X_{n+1}和贝叶斯信度估计值均方误差最小的线性估计形式。本节将在一些特殊的假定下，推导出线性估计的具体形式。我们将会发现，这种线性估计形式最终可以写为经验估计和先验估计的信度加权形式。这个信度因子称为 **Bühlmann 信度因子**。由于 Bühlmann 信度估计是均方误差最小的估计，因此也被称为**最大精度信度模型**，或**一致最精确信度**。

沿用上节的记号和假设。假设每一份保单对应一个风险参数，风险参数是一个随机变量 Θ，而且 Θ 的取值是无法观察的，但可以知道风险参数 Θ 的先验分布。损失 X 是一个随机变量，对于 Θ 的每一个具体值 θ，对应一个 X 的条件分布 $f_{X|\Theta}(x|\theta)$。对于一个给定的 θ，每次损失的条件分布 $X_i|\theta$ 是独立的。

12.4.1 Bühlmann 模型

Bühlmann 信度是最早使用的，也是最简单的最大精度信度模型。它假设对于给定的风险参数，每次损失的条件分布是同分布的。因此每次损失的条件均值和条件方差是相同的，用数学符号表示为：

$$\mu(\theta) = E(X_i|\Theta = \theta); \ v(\theta) = Var(X_i|\Theta = \theta)$$

注意 $\mu(\theta)$和 $v(\theta)$随 θ 变化，因此可以看做 θ 的函数。因为 Θ 是随机变量，所以 $\mu(\Theta)$和 $v(\Theta)$也是随机变量，我们可以对它们取均值和方差，记：

$$\mu = E[\mu(\Theta)], \ v = E[v(\Theta)], \ a = Var[\mu(\Theta)] \quad (12.4.1)$$

式（12.4.1）中的 μ 称为条件期望的期望。根据条件期望迭代法则，

可证明 $\mu = E(X_i)$，是没有任何关于 θ 的信息下的所用的估计量。式（12.4.1）中的 v 称为条件方差的期望，也称组间方差的期望；a 称为条件期望的方差，也称各组均值的方差。根据方差分解公式可知：

$$Var(X_i) = Var[E(X_i \mid \Theta)] + E[Var(X_i \mid \Theta)] = a + v \tag{12.4.2}$$

从式（12.4.2）知，X 的方差可以分为两部分：组间方差的期望 v 和各组均值的方差 a。

$$\begin{aligned} Cov(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\ &= E[E(X_i X_j \mid \Theta)] - E[E(X_i \mid \Theta)]E[E(X_j \mid \Theta)] \end{aligned}$$

利用 $X_i \mid \theta$ 是条件独立的，上式等于

$$\begin{aligned} & E[E(X_i \mid \Theta)E(X_j \mid \Theta)] - E[E(X_i \mid \Theta)]\ E[E(X_j \mid \Theta)] \\ &= E[\mu^2(\Theta)] - E^2[\mu(\Theta)] \\ &= Var[\mu(\Theta)] = a \end{aligned} \tag{12.4.3}$$

根据上面的分析，我们记 $E(X_i) = \mu, Var(X_i) = a + v, Cov(X_i, X_j) = a$，下面计算 Bühlmann 信度估计的具体形式。由无偏方程（12.3.13）得到：

$$\mu = \hat{\alpha}_0 + \mu \sum_{j=1}^{n} \hat{\alpha}_j \tag{12.4.4}$$

$$\sum_{j=1}^{n} \hat{\alpha}_j = 1 - \hat{\alpha}_0 / \mu \tag{12.4.5}$$

另外，式（12.3.15）的 n 个方程变为：

$$a = \sum_{\substack{j=1 \\ j \neq i}}^{n} \hat{\alpha}_i a + \hat{\alpha}_i (a + v) = \sum_{j=1}^{n} \hat{\alpha}_j a + \hat{\alpha}_i v, \qquad i = 1, \cdots, n$$

于是，

$$\hat{\alpha}_i = \frac{a(1 - \sum_{j=1}^{n} \hat{\alpha}_j)}{v} = \frac{a\hat{\alpha}_0}{\mu v}$$

将 i 从 1 到 n 相加得到：

$$\sum_{i=1}^{n} \hat{\alpha}_i = \sum_{j=1}^{n} \hat{\alpha}_j = \frac{na\hat{\alpha}_0}{\mu v}$$

结合方程（12.4.5），我们有

$$1 - \hat{\alpha}_0 / \mu = \frac{na\ \hat{\alpha}_0}{\mu v}$$

解出 $\hat{\alpha}_0$ 为：

$$\hat{\alpha}_0 = \frac{v\mu}{v + na}$$

从而得到：

$$\hat{\alpha}_j = \frac{a\ \hat{\alpha}_0}{\mu v} = \frac{a}{v + na}$$

因此，X_{n+1} 信度估计为：

$$\hat{\alpha}_0 + \sum_{j=1}^{n} \hat{\alpha}_j X_j = \frac{v}{v + na}\mu + \frac{na}{v + na}\bar{X} = (1 - z)\mu + z\bar{X} \quad (12.4.6)$$

其中，$z = \frac{na}{v + na} = \frac{n}{n + k}$，$k = \frac{v}{a} = \frac{E[Var(X_j \mid \Theta)]}{Var[E(X_j \mid \Theta)]}$

这里的 $z = \frac{n}{n + k} = \frac{n}{n + v/a}$ 为 Bühlmann 信度因子。

从式（12.4.6）可以看出，在 X_i 关于 θ 条件独立同分布的假定下，Bühlmann 信度估计可以写为经验保费和先验保费的信度加权形式。这种形式从直观上解释也是合理的。由于 X 的方差可以分为组间方差的期望 v 和各组方均值的方差 a 两部分。若保单组合相对于风险参数 Θ 来说具有风险同质性，这时 $\mu(\Theta) = E(X_j \mid \Theta)$ 的波动性较小，也就是说各组间的均值比较接近而各组的经验数据波动较大，这时 a 的值相对于 v 来说偏小，$k = v/a$ 将偏大，z 接近于 0，我们更加依赖于先验估计 μ 来厘定保费。当保单组合具有风险异质性，或 $\mu(\Theta) = E(X_j \mid \Theta)$ 的波动性较大，也就是说各组间的均值波动较大而各组内的经验数据波动较小，即 a 的值相对于 v 来说较大，则 k 的值较小，z 将接近于 1，这时我们将主要依赖样本均值 $\bar{X}$ 来厘定保费。因此 Bühlmann 信度估计不仅继承了信度保费的最小偏差的优良性质，与现实比较接近，而且所需要的参数都容易根据模型分布和先验分布求一阶矩和二阶矩得到，所以它的应用非常广泛。

【例 12－14】 承例 12－11，假设某保险公司承保的业务中，有 1/3 的保单的每月损失次数服从均值为 4 的泊松分布，有 2/3 的保单每月损失次数服从均值为 2 的泊松分布。对于一个随机抽取的保单，已知同一张保单每期的损失是独立的，观察到上个月的损失次数为 1。求这个月损失的 Bühlmann 信度估计。

解： 根据题目信息，求得

$$\mu = E[\mu(\Theta)] = \frac{1}{3} \times 4 + \frac{2}{3} \times 2 = \frac{8}{3}$$

$$v = E[v(\Theta)] = \frac{1}{3} \times 4 + \frac{2}{3} \times 2 = \frac{8}{3}$$

$$a = Var[\mu(\Theta)] = \frac{1}{3} \times 16 + \frac{2}{3} \times 4 - \left(\frac{8}{3}\right)^2 = \frac{8}{9}$$

继而我们可以求出 Bühlmann 信度因子 $z = \frac{n}{n + v/a} = \frac{1}{1 + 3} = 0.25$

因此，Bühlmann 估计值为 $(1 - z)\mu + z\bar{X} = 0.75 \times \frac{8}{3} + 0.25 \times 1 = 2.25$ ■

我们在例 12.10 中求得的贝叶斯信度估计值为 2.24，两者非常接近。

【例 12－15】 设 $\{X_j \mid \Theta;\ j = 1, \cdots, n + 1\}$ 是独立同分布的泊松随机变量，参数为 Θ，而 Θ 服从伽玛分布：

$$\pi(\theta)=\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}},\theta>0$$

求 X_{n+1}的 Bühlmann 信度估计。

解： 由于 $\{X_j\mid\Theta;\ j=1,\cdots,n+1\}$ 服从泊松分布，

$$\mu(\theta)=E(X_j\mid\Theta=\theta)=\theta;\ v(\theta)=Var(X_j\mid\Theta=\theta)=\theta$$

因此，

$$\mu=E[\mu(\Theta)]=E(\Theta)=\alpha\beta;\ v=E[v(\Theta)]=E(\Theta)=\alpha\beta$$

$$a=Var[\mu(\Theta)]=Var(\Theta)=\alpha\beta^2$$

于是，$k=\dfrac{v}{a}=\dfrac{\alpha\beta}{\alpha\beta^2}=\dfrac{1}{\beta}$

$$z=\frac{n}{n+k}=\frac{n}{n+1/\beta}=\frac{n\beta}{n\beta+1}$$

X_{n+1}的 Bühlmann 信度估计为：

$$z\bar{X}+(1-z)\mu=\frac{n\beta}{n\beta+1}\bar{X}+\frac{1}{n\beta+1}\alpha\beta$$ ■

与例 12 - 13 进行对比可以看出，在例 12 - 13 中，当 $m_j=1$，$r=0$，$c=1$ 时，情况与本题一致，贝叶斯信度估计值为：

$$E(X_{n+1}\mid\vec{X})=\alpha_{.}\beta_{.}=\frac{\alpha+n\bar{X}}{1/\beta+n}=\frac{n\beta}{n\beta+1}\bar{X}+\frac{1}{n\beta+1}\alpha\beta$$

与本例中的结果是一致的。当 Bühlmann 信度估计与贝叶斯信度估计值完全相等时，我们称之为**精确信度**。当贝叶斯信度估计值自身为经验数据的线性组合时，则达到精确信度。事实上，当 $\{X_j,\ j=1,\cdots,n\}$ 的条件分布和先验分布满足共轭分布的条件时，可以证明达到精确信度。此时，

$$f_{X_j\mid\Theta}(x_j\mid\theta)=\frac{p(x_j)e^{-\theta x_j}}{q(\theta)},\quad \pi(\theta)=\frac{[q(\theta)]^{-k}e^{-\theta\mu k}}{c(\mu,k)}$$

其中 θ 取有限实值，μ，k 正是信度保费模型中的参数。有兴趣的读者可自己验证上述结论。

12.4.2　Bühlmann - Straub 模型

Bühlmann 模型假定 $\{X_j\mid\Theta;j=1,\cdots,n\}$ 是独立同分布的随机变量，但在许多情况下，这个假定不一定得到满足。例如，在团体保险模型中，单个被保险人的风险水平在各年没有变化，但团体的人数每年都不同，我们只知道此群体每年的总索赔额和人数，此时，由于被保险人数的变化，该团体每年的人均索赔额的均值虽然相同，但是方差却不同了。又如，如果某被保险人某一年的索赔次数数据记录的月份不全，我们只知道每年记录的总次数和记录月份，而不知道每个月的次数，也会导致每年索赔次数条件独立同分布的假设不成立。这时该如何估计该团体（被保险人）的明年

总索赔额（次数）的期望值呢？

注意上述情况有一个共同特点，虽然我们不知道此群体单位时间每人的损失数据，但是一个群体或一段时间内的损失数据的总值或均值容易得到。假设$\{X_j \mid \Theta; j=1,\cdots,n\}$为群体内人均损失或单位时间的平均损失。若该群体内每个被保险人的风险参数θ是相同的，则$\{X_j \mid \Theta; j=1,\cdots,n\}$条件均值不变，条件方差改变。可考虑下面的 Bühlmann - Straub 模型来解决这类问题。假设在给定$\Theta=\theta$条件下，$X_1,\cdots,X_n$是相互条件独立的，具有相同的条件均值

$$\mu(\theta)=E(X_j \mid \Theta=\theta)$$

但条件方差不相同：

$$Var(X_j \mid \Theta=\theta)=v(\theta)/m_j$$

其中m_j是一个已知的常数。在团体保险中，m_j可以看做团体在第j年的总风险暴露数或者第j年的被保险人数。在给定$\Theta=\theta$条件下，X_j可以看做第j年m_j个独立同分布被保险人的人均索赔额。与 Bühlmann 模型一样，定义

$$\mu=E[\mu(\Theta)],\quad v=E[v(\Theta)],\quad a=Var[\mu(\Theta)]$$

由式（12.4.2）和式（12.4.3），可以得到$E(X_j)=\mu$和$Cov(X_i,X_j)=a$，以及

$$\begin{aligned}Var(X_j)&=E[Var(X_j \mid \Theta)]+Var[E(X_j \mid \Theta)]\\&=E[v(\Theta)/m_j]+Var[\mu(\Theta)]\\&=v/m_j+a\end{aligned}\tag{12.4.7}$$

下面计算X_{n+1}的信度估计。为方便起见，令$m=m_1+\cdots+m_n$。由无偏方程式（12.3.13）得到：

$$\sum_{j=1}^{n}\hat{\alpha}_j=1-\hat{\alpha}_0/\mu\tag{12.4.8}$$

对于$i=1,\cdots,n$，式（12.3.15）变为：

$$a=\sum_{\substack{j=1\\j\neq i}}^{n}\hat{\alpha}_j a+\hat{\alpha}_i(a+v/m_i)=\sum_{j=1}^{n}\hat{\alpha}_j a+v\hat{\alpha}_i/m_i$$

即

$$\hat{\alpha}_i=\frac{a}{v}m_i\left(1-\sum_{j=1}^{n}\hat{\alpha}_j\right)=\frac{a}{v}\frac{\hat{\alpha}_0}{\mu}m_i,\qquad i=1,\cdots,n\tag{12.4.9}$$

于是利用式（12.4.8）和（12.4.9）得到：

$$1-\frac{\hat{\alpha}_0}{\mu}=\sum_{j=1}^{n}\hat{\alpha}_j=\sum_{i=1}^{n}\hat{\alpha}_i=\frac{a}{v}\frac{\hat{\alpha}_0}{\mu}\sum_{i=1}^{n}m_i=\frac{a}{v}\frac{\hat{\alpha}_0}{\mu}m$$

由此解得：

$$\hat{\alpha}_0=\frac{\mu}{1+\dfrac{am}{v}}=\frac{v/a}{m+v/a}\mu$$

$$\hat{\alpha}_j=\frac{a\,\hat{\alpha}_0}{\mu v}\cdot m_j=\frac{m_j}{m+v/a}$$

同样，信度估计可以写为：

$$\hat{\alpha}_0 + \sum_{j=1}^{n} \hat{\alpha}_j X_j = z\bar{X} + (1-z)\mu \tag{12.4.10}$$

其中，

$$z = \frac{m}{m+k}, k = \frac{v}{a}, \bar{X} = \sum_{j=1}^{n} \frac{m_j}{m} X_j$$

事实上，m_jX_j 可以看做团体保险中团体在第 j 年的总损失，m 为 n 年内的总被保险人个数，因此 $\bar{X}$ 实际上是这个团体在 n 年内的人均损失。从这个意义上说，那么 Bühlmann – Straub 信度保费的形式与 Bühlmann 模型是一样的。

【例 12 –16】 假设被保险人可分为若干个风险子集，每个风险子集内的每个被保险人具有相同的风险参数 Θ。被保险人在一年内索赔次数服从二项分布 $B(3, \Theta)$，Θ 是随机变量，其分布密度为 $\pi(\theta)=6\theta(1-\theta)$，$0<\theta<1$。已知在 1997 年每个风险子集有 10 个被保险人在某保单组合中；1998 年每个风险子集有 12 个被保险人，1999 年有每个风险子集有 15 个被保险人。随机记录一个风险子集，它在这 3 年的总索赔次数为 18（1997 年），20（1998 年）和 27（1999 年）。假设 2000 年该风险子集共有 20 个被保险人在保单组合中，使用 Bühlmann – Straub 模型求 2000 年该风险子集的总索赔次数的信度估计。

解： 设 W 表示被保险人在一年内的索赔次数，则

$$\mu(\theta) = E(W \mid \Theta=\theta) = 3\theta$$

$$v(\theta) = Var(W \mid \Theta=\theta) = 3\theta(1-\theta)$$

又由 Θ 的分布，可以计算出：

$$\mu = E[\mu(\Theta)] = E[3\Theta] = \int_0^1 3\theta \cdot 6\theta(1-\theta)d\theta = \frac{3}{2}$$

$$v = E[v(\Theta)] = E[3\Theta(1-\Theta)] = \int_0^1 3\theta(1-\theta)\cdot 6\theta(1-\theta)d\theta = \frac{3}{5}$$

$$a = Var[\mu(\Theta)] = Var[3(\Theta)] = 9Var(\Theta)$$

$$= 9[\int_0^1 \theta^2 6\theta(1-\theta)d\theta - \left(\frac{1}{2}\right)^2] = \frac{9}{20}$$

$$m = m_1 + m_2 + m_3 = 10 + 12 + 15 = 37$$

$$k = \frac{v}{a} = \frac{4}{3}$$

随机选取一个风险子集，由于在 1997 年每个风险子集有 10 个被保险人，则该子集的人均索赔次数为：

$$X_1 = \frac{W_{1,1} + \cdots + W_{1,10}}{10} = \frac{18}{10},\quad m_1 = 10$$

相应地，在 1998 年和 1999 年有：

$$X_2=\frac{20}{12},\ m_2=12 \qquad X_3=\frac{27}{15},\ m_3=15$$

应用 Bühlmann－Straub 模型，信度因子为：

$$z=\frac{m}{m+k}=\frac{37}{37+4/3}=0.965$$

$$\bar{X}=\sum_{i=1}^{n}\frac{m_i}{m}X_i=\frac{10\times\frac{18}{10}+12\times\frac{20}{12}+\frac{27}{15}\times 15}{37}=1.757$$

因此，2000 年该风险子集的人均索赔次数的信度估计为：

$$z\bar{X}+(1-z)\mu=0.965\times 1.757+0.035\times 1.5=1.748$$

该风险子集的总索赔次数的信度估计为 $20\times 1.748=34.96$。■

12.4.3 Bühlmann－Straub 模型的推广

在团体保险中，Bühlmann－Straub 模型假定风险子集内的每个被保险人的风险水平都是独立同分布的。这个假定在保险实际中并不一定满足。实际上，绝对独立同质的风险并不存在。当某群体中的所有个体并非完全独立时，群体均值的方差就会大于个体方差除以个体数。下面的例子给出这种情况下 Bühlmann－Straub 模型的推广。

【例 12－17】 设在给定 Θ 的条件下，$\{X_j, j=1,\cdots,n\}$ 是相互独立的随机变量，$E(X_j|\Theta)=\mu(\Theta)$，$Var(X_j|\Theta)=w(\Theta)+v(\Theta)/m_j$，求 X_{n+1} 的信度估计。

解： 根据条件可知，

$$\begin{aligned}E(X_j)&=E[E(X_j|\Theta)]=E[\mu(\Theta)]=\mu\\ Var(X_j)&=E[Var(X_j|\Theta)]+Var[E(X_j|\Theta)]\\ &=E[w(\Theta)+v(\Theta)/m_j]+Var[\mu(\Theta)]\\ &=w+v/m_j+a\end{aligned}$$

对于 $i\neq j$，$Cov(X_i,X_j)=a$，将上式代入式（12.3.14）得到：

$$a=\sum_{j=1}^{n}\hat{\alpha}_j a+\hat{\alpha}_i(w+v/m_i)=a(1-\hat{\alpha}_0/\mu)+\hat{\alpha}_i(w+v/m_i),i=1,\cdots,n$$

因此，$\hat{\alpha}_i=\dfrac{a\hat{\alpha}_0/\mu}{w+v/m_i}$。

将 i 相加得到：

$$\frac{a\hat{\alpha}_0}{\mu}\sum_{j=1}^{n}\frac{m_j}{v+wm_j}=\sum_{j=1}^{n}\hat{\alpha}_j=1-\hat{\alpha}_0/\mu$$

因此，$\hat{\alpha}_0=\dfrac{1}{\dfrac{a}{\mu}\sum_{j=1}^{n}\dfrac{m_j}{v+wm_j}+\dfrac{1}{\mu}}=\dfrac{\mu}{1+am^*}$

其中，$m^{*}=\sum_{j=1}^{n}\frac{m_j}{v+wm_j}$，于是，

$$\hat{\alpha}_j=\frac{am_j}{v+wm_j}\frac{1}{1+am^{*}}$$

信度估计值为：

$$\frac{\mu}{1+am^{*}}+\frac{a}{1+am^{*}}\sum_{j=1}^{n}\frac{m_jX_j}{v+wm_j}$$

若令 $\bar{X}=\frac{1}{m^{*}}\sum_{j=1}^{n}\frac{m_j}{v+wm_j}X_j,z=\frac{am^{*}}{1+am^{*}}$

信度估计值等于 $z\bar{X}+(1-z)\mu$。注意，若令 m_j 趋于无穷，则 $z\longrightarrow\frac{an/w}{1+an/w}<1$，而在 Bühlmann - Straub 中 z 是趋于 1 的。这说明无论风险暴露数有多大，信度因子总小于 1，经验数据总不是完全可信的。■

§12.5　经验贝叶斯信度参数估计

在前面几节介绍的信度模型中，都需要先假设先验分布和条件分布已知，然后得到参数 $\mu=E[E(X_{ij}\mid\Theta_i)]$，$v=E[Var(X_{ij}\mid\Theta_i)]$，和 $a=Var[E(X_{ij}\mid\Theta_i)]$，进而得到信度估计值。然而在实际中，比如设计新险种时，先验信息并不充足，我们不是总知道先验分布和损失分布，或者只能合理假设分布的形式，而不知道分布的具体参数。我们常常需要根据具体数据和先验知识进行参数估计，这种方法就叫做经验贝叶斯估计。

根据已知信息的多少，经验贝叶斯估计可以分为非参数方法、半参数方法和参数方法。如果对 $\pi(\theta)$ 和 $f_{X|\Theta}(x\mid\theta)$ 的分布形式均没有先验知识的情况，称为**非参数情形**；如果不知道 $\pi(\theta)$ 的分布形式但可以对 $f_{X|\Theta}(x\mid\theta)$ 作出某种参数形式分布的假设时，则属于半参数情形；如果对 $\pi(\theta)$ 和 $f_{X|\Theta}(x\mid\theta)$ 均能作出具有参数形式的假设，需要估计未知参数，则属于**参数情形**。

在这一节中数据将以如下形式表示：对 $r\geq1$ 个被保险人，其中每个人的单位风险损失量是 $\mathrm{X}_i=(X_{i1},\cdots,X_{in_i})^T$，$i=1,\cdots,r$。随机向量 $\{\mathrm{X}_i,\ i=1,\cdots,r\}$ 在统计上认为是相互独立的（即假设不同投保人的经验数据相互独立）。第 i 个被保险人的未知风险参数是 θ_i，$i=1,\cdots,r$，并进一步假定 θ_i 是独立同分布的，具有结构密度 $\pi(\theta_i)$ 的随机变量 Θ_i 的一次实现（$i=1,\cdots,r$）。对固定的 i，假设条件随机变量 $X_{ij}\mid\Theta_i$ 相互独立，有概率密度 $f_{X_{ij}\mid\Theta}(x_{ij}\mid\theta)$，$j=1,\cdots,n_i$。

有两个常见的情形可以产生上述的数据形式。第一种情形是分类费率厘定。下标 i 表示类别或团体，j 表示当中的个体。第二种情形类似，下标 i 仍指代类别或团体，j 表示年份，以每年的平均损失量作为观察值。

例如，10 个不同风险类别的被保险人，观察 $j=1$，2，3 年的理赔额数据。

经验数据还应当包括被保险人 i 的风险量向量 $m_i=(m_{i1},\cdots,m_{in_i})^T$，$i=1,\cdots,r$。为了表达上的方便，记

$$m_i=\sum_{j=1}^{n_i}m_{ij},\quad m=\sum_{i=1}^{r}m_i$$

为被保险人 i 的过去总风险量。记

$$\bar{X}_i=\frac{1}{m_i}\sum_{j=1}^{n_i}m_{ij}X_{ij}$$

为被保险人 i 的过去单位风险平均损失。另外所有被保险人单位风险平均损失为：

$$\bar{X}=\frac{1}{m}\sum_{i=1}^{r}m_i\bar{X}_i$$

$$=\frac{1}{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i}m_{ij}X_{ij}$$

对于每个被保险人 i，在给定风险参数 $\Theta_i=\theta_i$ 的情况下，$X_{i1}\mid\theta_i,X_{i2}\mid\theta_i$，$\cdots X_{in_i}\mid\theta_i$ 是独立的。此时平均风险损失量有条件均值和条件方差：

$$E[X_{ij}\mid\Theta_i=\theta_i]=\mu(\theta_i),\quad Var[X_{ij}\mid\Theta_i=\theta_i]=\frac{v(\theta_i)}{m_{ij}}$$

由于每个投保人的风险参数是同分布的，所以结构参数可以记为 $\mu=E[\mu(\Theta_i)]$，$v=E[v(\Theta_i)],a=Var[\mu(\Theta_i)]$。

在得到了风险参数 μ，v 和 a 的值后，我们可以计算出各个被保险人的信度因子 Z_i 和信用保费。经验贝叶斯方法可以应用于 Bühlmann 模型和 Bühlmann – Straub 模型。需要注意的是，即使我们得到的估计值是结构参数的无偏估计，并不能代表进一步估计而得的信度因子和信度保费就是无偏的。比如即使 $E[\hat{v}]=v$，$E[\hat{a}]=a$，并不一定意味着 $\hat{k}=\hat{v}/\hat{a}$ 就是 $k=v/a$ 的无偏估计。

12.5.1 非参数估计

在非参数情形下，我们对 $\pi(\theta)$ 和 $f_{X|\Theta}(x\mid\theta)$ 的分布均无法作出具有参数情形的假设。此时，为了估计下一年的信度保费，我们需要利用经验数据，得到未知参数 μ、v 和 a 的估计值 $\hat{\mu}$、$\hat{v}$ 和 $\hat{a}$，则被保险人下一年的信度保费的估计值为：

$$\hat{Z}_i\bar{X}_i+(1-\hat{Z}_i)\hat{\mu},\qquad i=1,2,\cdots,r \tag{12.5.1}$$

其中，$\hat{Z}_i=\dfrac{m_i}{m_i+\hat{k}}$，$\hat{k}=\dfrac{\hat{v}}{\hat{a}}$

下面分别介绍非参数经验贝叶斯估计应用于 Bühlmann 模型和 Bühlmann-Straub 模型的情形。

1. Bühlmann 模型。在 Bühlmann 模型中，每个被保险人的“风险期间”是相同的，即 $n_1 = n_2 = \cdots n_r = n$。并且每个“风险期间”的风险量为 1，即 $m_{ij} = 1$，$i = 1, 2, \cdots, r$，$j = 1, 2, \cdots, n$。这可以理解为 r 个被保险人，每人都有 n 年的观察数据，且每年只观察一次。

考虑第 i 个被保险人的平均损失：

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \tag{12.5.2}$$

所有被保险人的均值为：

$$\bar{X} = \frac{1}{r}\sum_{i=1}^{r} \bar{X}_i = \frac{1}{rn}\sum_{i=1}^{r}\sum_{j=1}^{n} X_{ij} \tag{12.5.3}$$

根据条件期望和方差公式：

$$E[X_{ij}] = E[E(X_{ij} \mid \Theta_i)] = E[\mu(\Theta_i)] = \mu \tag{12.5.4}$$

$$Var[\bar{X}_i \mid \Theta_i = \theta_i] = \frac{v(\theta_i)}{n} \tag{12.5.5}$$

$$E[\bar{X}] = E[\bar{X}_i] = \mu \tag{12.5.6}$$

$$Var[\bar{X}_i] = E[Var(\bar{X}_i \mid \Theta_i)] + Var[E(\bar{X}_i \mid \Theta_i)] = \frac{v}{n} + a \tag{12.5.7}$$

下面我们介绍结构参数 μ、v 和 a 的估计。对于 μ，考虑估计量 $\hat{\mu} = \bar{X}$。由式（12.5.3）有

$$\begin{aligned} E(\bar{X}) &= (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n} E(X_{ij}) = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n} E(E(X_{ij} \mid \Theta_i)) \\ &= (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n} E(\mu(\Theta_i)) = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mu = \mu \end{aligned}$$

因此 $\hat{\mu} = \bar{X}$ 为 μ 的一个无偏估计。

对于 v，首先考虑 $\hat{v}_i = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$。由于给定 $\Theta_i = \theta_i$，X_{ij}，$j = 1, \cdots, n$ 是相互独立的，因此给定 $\Theta_i = \theta_i$，$\hat{v}_i$ 是 $v(\theta_i)$ 的无偏估计，且

$$E(\hat{v}_i) = E[E(\hat{v}_i \mid \Theta_i)] = E[v(\Theta_i)] = v \tag{12.5.8}$$

即 $\hat{v}_i$ 是 v 的无偏估计。于是统计量

$$\hat{v} = \frac{1}{r}\sum_{i=1}^{r}\hat{v}_i = \frac{1}{r(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2 \tag{12.5.9}$$

也是 v 的无偏估计。

接下来估计参数 a，由式(12.5.6)和(12.5.7)知，$\bar{X}_1$，$\bar{X}_2, \cdots, \bar{X}_r$ 相互独立，且有相同的期望 μ 和方差 $a + v/n$。考虑估计量

$$\hat{a} = \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2 - \frac{\hat{v}}{n} \tag{12.5.10}$$

由式（12.5.7）知：

$$\begin{aligned} E(\hat{a}) &= \frac{1}{r-1}E\left[\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2\right] - E\left[\frac{\hat{v}}{n}\right] \\ &= \frac{1}{r-1}E\left[\sum_{i=1}^{r}(\bar{X}_i - \mu)^2 - r(\mu - \bar{X})^2\right] - \frac{v}{n} \\ &= \frac{\sum_{i=1}^{r} Var[\bar{X}_i] - rVar[\bar{X}]}{r-1} - \frac{v}{n} \\ &= \frac{r-1}{r-1}\left(\frac{v}{n} + a\right) - \frac{v}{n} = a \end{aligned} \tag{12.5.11}$$

因此，$\hat{a} = \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2 - \frac{\hat{v}}{n}$ 是 a 的一个无偏估计。

式（12.5.1）可以用单因素方差分析来解释。事实上，我们可以把拥有不同的风险参数的不同被保险人看做是方差分析中 r 种不用的处理方式，同一被保险人的 n 个损失数据可以看做是对同一处理方式的 n 个观察值。那么 v 考察的就是同一种处理方式的不同数据之间差距的平均水平。而 a 则考察不同处理方式平均水平之间的差距。而 $\hat{v}$ 恰恰为组内误差，而 $\hat{a}$ 的前一项等于组间方差除以 n，这一项不能完全反映组间差异，因为它包含了样本的一部分随机性，因此减去 $\hat{a}$ 中的后一项，得到了组间差距无偏估计。当组间平方和相对组内平方和而言比较小时——也就是 $\hat{a}$ 相对 $\hat{v}$ 比较小时，接受所有处理方法有相同均值的原假设。这也意味着 $\hat{Z}$ 接近于 0，即给予每个 $\bar{X}_i$ 很小的信度，这在所有被保险人风险同质的情形下是一个很自然的结论。

由于估计值中存在减号，因此可能会出现 $\hat{a}<0$ 的情况，这代表了组间差距非常小，此时我们令 $\hat{Z}=0$。这种情形等价于在方差分析中 F 统计量小于 1，此时我们无法拒绝均值相等的原假设。

【例 12－18】 假设有两个被保险人，他们过去三年的损失额数据如表 12－2所示。

表 12－2

被保险人	保单年度		
	1	2	3
1	4	8	9
2	11	13	9

（1）估计两个被保险人在下一年的 Bühlmann 信度估计；

（2）如果第二个被保险人的经验数据改为 $\bar{x}_2=(5, 6, 10)$，分别估计

两位被保险人下一年的信度估计。

解：(1) 由表 12－2 知，先 $r=2$，$n_1=n_2=n=3$，$m_{ij}=1$，$m_1=m_2=3$，此数据适合应用 Bühlmann 模型。

根据数据，可以求出：

$$\bar{X}_1=\frac{4+8+9}{3}=7,\quad \bar{X}_2=\frac{11+13+9}{3}=11,\quad 以及\ \bar{X}=\frac{7+11}{2}=9=\hat{\mu}$$

根据式 (12.5.9) 和式 (12.5.10)，分别求出：

$$\begin{aligned}\hat{v}&=\frac{1}{r(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2\\&=\frac{1}{2\times2}[(4-7)^2+(8-7)^2+(9-7)^2+(11-11)^2\\&\quad+(13-11)^2+(9-11)^2]=\frac{11}{2}\end{aligned}$$

$$\begin{aligned}\hat{a}&=\frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2-\frac{\hat{v}}{n}\\&=\frac{1}{1}\times[(7-9)^2+(11-9)^2]-\frac{11}{3}=\frac{13}{3}\end{aligned}$$

进而，

$$\hat{k}=\frac{\hat{v}}{\hat{a}}=\frac{11/2}{13/3}=1.27,\qquad \hat{Z}_1=\frac{m_1}{m_1+\hat{k}}=\frac{3}{3+1.27}=0.71=\hat{Z}_2$$

两位被保险人下一年的信度保费估计分别为：

$$\hat{Z}_1\bar{X}_1+(1-\hat{Z}_1)\hat{\mu}=0.71\times7+(1-0.71)\times9=7.59$$

$$\hat{Z}_2\bar{X}_2+(1-\hat{Z}_2)\hat{\mu}=0.71\times11+(1-0.71)\times9=10.41$$

(2) 此时，$\bar{X}_1=\frac{4+8+9}{3}=7$，$\bar{X}_2=\frac{5+6+10}{3}=7$，以及 $\bar{X}=\frac{7+7}{2}=7=\hat{\mu}$。

根据式 (12.5.9) 和式 (12.5.10)，分别求出：

$$\begin{aligned}\hat{v}&=\frac{1}{r(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2\\&=\frac{1}{2\times2}[(4-7)^2+(8-7)^2+(9-7)^2+(5-7)^2\\&\quad+(6-7)^2+(10-7)^2]=7\end{aligned}$$

$$\begin{aligned}\hat{a}&=\frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2-\frac{\hat{v}}{n}\\&=\frac{1}{1}\times[(7-7)^2+(7-7)^2]-\frac{7}{3}=-\frac{7}{3}\end{aligned}$$

出现 $\hat{a}<0$ 的情况，此时采用 $\hat{Z}=0$。所以两位被保险人在下一年的信度保费估计值均为 $\hat{\mu}=7$。 ■

2. Bühlmann－Straub 模型。该模型考虑了不同投保团体中个体数或者

观测年数不同的情况，因此数据背景与本节开始介绍的是一致的：对于 r 个投保团体，其中第 i 个被保险人（$i=1,2,\cdots,r$）的"风险期间"数为 n_i，其中第 i 个团体的第 j 个"风险期间"有 m_{ij} 个风险量，单位风险损失量为 X_{ij}。其中单位风险损失量 X_{ij} 是随机变量，第 i 个被保险人的单位风险损失量记为 $\vec{X}_i=(X_{i1},\ X_{i2},\ \cdots X_{in_i})$，不同被保险人之间的单位风险损失量是独立的。由于

$$E[X_{ij}]=E[E(X_{ij}\mid\Theta_i)]=E[\mu(\Theta_i)]=\mu$$

$$Var[X_{ij}\mid\Theta_i]=\frac{v(\Theta_i)}{m_{ij}}$$

$$Var[X_{ij}]=Var[E(X_{ij}\mid\Theta_i)]+E[Var(X_{ij}\mid\Theta_i)]=a+\frac{v}{m_{ij}}$$

由于 X_{i1}，X_{i2}，$\cdots X_{in_i}$ 关于 Θ_i 的条件独立，则

$$E[\bar{X}_i\mid\Theta_i]=\frac{1}{m_i}\sum_{j=1}^{n_i}m_{ij}E[X_{ij}\mid\Theta_i]=\mu(\Theta_i),E[\bar{X}_i]=\mu \quad (12.5.12)$$

$$Var[\bar{X}_i\mid\Theta_i]=\frac{1}{m_i^2}\sum_{j=1}^{n_i}m_{ij}^2Var[X_{ij}\mid\Theta_i]=\frac{1}{m_i^2}\sum_{j=1}^{n_i}m_{ij}v(\Theta_i)=\frac{v(\Theta_i)}{m_i}$$

$$Var[\bar{X}_i]=E[Var(\bar{X}_i\mid\Theta_i)]+Var[E(\bar{X}_i\mid\Theta_i)]=\frac{v}{m_i}+a \quad (12.5.13)$$

$$E[\bar{X}]=\frac{1}{m}\sum_{i=1}^{r}m_iE[\bar{X}_i]=\mu$$

由于 $\bar{X}_1$，$\bar{X}_2,\cdots,\bar{X}_r$ 是独立的，所以，

$$Var[\bar{X}]=\frac{1}{m^2}\sum_{i=1}^{r}m_i^2Var[\bar{X}_i]=\frac{v}{m}+\frac{\sum_{i=1}^{r}m_i^2}{m^2}a \quad (12.5.14)$$

下面我们给出结构参数 μ，v 和 a 的估计。

根据式（12.5.12），μ 的一个无偏估计是 $\hat{\mu}=\bar{X}$。若记

$$\hat{v}_i=\frac{\sum_{j=1}^{r}m_{ij}(X_{ij}-\bar{X}_i)^2}{n_i-1}$$

则

$$E[\hat{v}_i]=\frac{E\{\sum_{j=1}^{n_i}m_{ij}[(X_{ij}-\bar{X}_i)^2]\}}{n_i-1}$$

$$=\frac{1}{n_i-1}E\{\sum_{j=1}^{n_i}m_{ij}[(X_{ij}-\mu+\mu-\bar{X}_i)^2]\}$$

$$=\frac{1}{n_i-1}E\{\sum_{j=1}^{n_i}m_{ij}[(X_{ij}-\mu)^2+2(X_{ij}-\mu)(\mu-\bar{X}_i)+(\mu-\bar{X}_i)^2]\}$$

$$= \frac{1}{n_i - 1} E\{ \sum_{j=1}^{n_i} m_{ij} [(X_{ij} - \mu)^2 - (\mu - \bar{X}_i)^2] \}$$

$$= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} m_{ij} [Var(X_{ij}) - Var(\bar{X}_i)]$$

$$= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} m_{ij} \left[\frac{v}{m_{ij}} - \frac{v}{m_i} \right] = v \tag{12.5.15}$$

因此，v_i，$i = 1, \cdots, r$ 都是 v 的无偏估计。这些无偏估计量的任意加权平均也是 v 的无偏估计。我们选择权重与 $n_i - 1$ 成比例，就是让原始的 X_{ij}的权重与 m_{ij}成比例，即取 $w_i = (n_i - 1) / \sum_{i=1}^{r} (n_i - 1)$，可以得到 v 的一个无偏估计量：

$$\hat{v} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} m_{ij} (X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^{r} (n_i - 1)} \tag{12.5.16}$$

下面估计 a，根据式（12.5.13）考虑估计量：

$$\hat{a} = [\sum_{i=1}^{r} m_i (\bar{X}_i - \bar{X})^2 - \hat{v}(r - 1)] / (m - m^{-1} \sum_{i=1}^{r} m_i^2) \tag{12.5.17}$$

由于

$$E[\sum_{i=1}^{r} m_i (\bar{X}_i - \bar{X})^2]$$

$$= E[\sum_{i=1}^{r} m_i (\bar{X}_i - \mu)^2] - E[m (\mu - \bar{X})^2]$$

$$= \sum_{i=1}^{r} m_i Var[\bar{X}_i] - m Var[\bar{X}]$$

$$= \sum_{i=1}^{r} m_i \left(\frac{v}{m_i} + a \right) - m \left(\frac{v}{m} + m^{-2} \sum_{i=1}^{r} m_i^2 a \right)$$

$$= (r - 1) v + \left(m - m^{-1} \sum_{i=1}^{r} m_i^2 \right) a$$

带入式（12.5.17），得到：

$$E[\hat{a}] = \{ E[\sum_{i=1}^{r} m_i (\bar{X}_i - \bar{X})^2] - (r - 1) E[\hat{v}] \} / (m - m^{-1} \sum_{i=1}^{r} m_i^2)$$

$$= (m - m^{-1} \sum_{i=1}^{r} m_i^2) a / (m - m^{-1} \sum_{i=1}^{r} m_i^2) = a \tag{12.5.18}$$

因此可见，$\hat{a} = [\sum_{i=1}^{r} m_i (\bar{X}_i - \bar{X})^2 - \hat{v}(r - 1)] / (m - m^{-1} \sum_{i=1}^{r} m_i^2)$ 为 a 的一个无偏估计。

与在 Bühlmann 模型中相似的，可能会出现 $\hat{a} < 0$ 的情况，此时经验数据表明不同的被保险人之间损失情况的差距很小，先验知识值得相信，取 $\hat{Z} = 0$。

值得注意的是，虽然所有参数的估计值都是无偏的，但这并不一定就

说明经验信度估计值也是无偏的。事实上我们可以求出所有被保险人的总损失 $TL = \sum_{i=1}^{r} m_i \bar{X}_i$。依照上述模型，如果在过去也按照上面计算出的信度保费收取保费，总保费应该是：

$$
\begin{aligned}
TP &= \sum_{i=1}^{r} m_i [\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i)\hat{\mu}] \\
&= \sum_{i=1}^{r} m_i \bar{X}_i - \sum_{i=1}^{r} m_i [(1 - \hat{Z}_i)\bar{X}_i - (1 - \hat{Z}_i)\hat{\mu}] \\
&= TL - \sum_{i=1}^{r} \frac{m_i \hat{k}}{m_i + \hat{k}} (\bar{X}_i - \hat{\mu})
\end{aligned}
$$

理想情况是 TL 等于 TP。因为任何增加保费的行为要想得到监管机构的同意，必须要以过去总的索赔水平为依据。信度保费需要有着理论和现实意义，同时如果还能够保持总保费收入与总损失相匹配就更好了。这应该有

$$
\sum_{i=1}^{r} \frac{m_i \hat{k}}{m_i + \hat{k}} (\bar{X}_i - \hat{\mu}) = 0
$$

即 $\sum_{i=1}^{r} \hat{Z}_i (\bar{X}_i - \hat{\mu}) = 0$。因此，

$$
\hat{\mu} = \frac{\sum_{i=1}^{r} \hat{Z}_i \bar{X}_i}{\sum_{i=1}^{r} \hat{Z}_i} \tag{12.5.19}
$$

式（12.5.19）的 μ 的估计是个体样本均值的信度加权平均，而非直接按风险量的加权平均。它的好处在于估计先验均值时考虑到了每个个体样本的可信度，使得总保费等于总损失。

【例 12－19】 假设两组被保险人在过去三年的经验数据和第四年的被保险人数如表 12－3 所示。

求：（1）对两组被保险人第四年应收取的信度保费。

（2）使用信度加权平均估计 μ，两组被保险人各自应缴纳的第四年的信度保费。

（3）如果第一组保险人的第二年总损失数据不是 18 000 而是 15 000，重新计算上面问题。

表 12－3

被保险人		保单年度			
		1	2	3	4
1	总损失	—	11 000	18 000	—
	团体人数	—	50	80	70
2	总损失	20 000	25 000	24 000	—
	团体人数	100	120	125	100

解：

(1) 根据数据，可以看出 $r=2$，$n_1=2$，$n_2=3$。对于第一组被保险人团体，下标表示哪一年对运算没有影响。方便起见，我们下标采取的年份和表中对应：

$m_{12}=50$，$X_{12}=11\ 000/50=220$，$m_{13}=80$，$X_{13}=18\ 000/80=225$，以及

$m_1=m_{12}+m_{13}=130$，$\bar{X}_1=(11\ 000+18\ 000)/130=223.08$。

对于第二组被保险团体，$m_{21}=100$，$X_{21}=20\ 000/100=200$，$m_{22}=120$，$X_{22}=25\ 000/120=625/3$，$m_{23}=125$，$X_{23}=24\ 000/125=192$，$m_2=m_{21}+m_{22}+m_{23}=345$，$m=m_1+m_2=475$，以及 $\bar{X}_2=\dfrac{20\ 000+25\ 000+24\ 000}{345}=200$

于是，$\hat{\mu}=\bar{X}=\dfrac{223.08\times130+200\times345}{130+345}=206.31$

根据式 (12.5.16) 和式 (12.5.17)，

$$\hat{v}=\frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}m_{ij}(X_{ij}-\bar{X}_i)^2}{\sum_{i=1}^{r}(n_i-1)}$$

$$=\frac{1}{1+2}[50\times(200-223.28)^2+\cdots+125\times(192-200)^2]$$

$$=5\ 700.855$$

$$\hat{a}=[\sum_{i=1}^{r}m_i(\bar{X}_i-\bar{X})^2-\hat{v}(r-1)]\Big/\left(m-\frac{\sum_{i=1}^{r}m_i^2}{m}\right)$$

$$=\frac{130\times(223.08-206.31)^2+345\times(200-206.31)^2-5\ 700.855\times1}{475-(130^2+345^2)/475}$$

$$=236.15$$

因此，$\hat{k}=\dfrac{\hat{v}}{\hat{a}}=24.15$，$\hat{Z}_1=\dfrac{m_1}{m_1+\hat{k}}=0.84$，$\hat{Z}_2=\dfrac{m_2}{m_2+\hat{k}}=0.93$

被保险人团体 1 和被保险人团体 2 的个体平均信度保费估计值分别为：

$$\hat{Z}_1\bar{X}_1+(1-\hat{Z}_1)\hat{\mu}=0.84\times223.08+(1-0.84)\times206.31=220.45$$

$$\hat{Z}_2\bar{X}_2+(1-\hat{Z}_2)\hat{\mu}=0.93\times200+(1-0.93)\times206.31=200.41$$

于是，两个被保险人团体下一年要缴纳的总的信度保费估计分别为 $220.45\times70=15\ 431.5$ 和 $200.41\times100=20\ 041$。

(2) 从上题中我们知道 $\hat{Z}_1=0.84$，$\hat{Z}_2=0.93$，根据式 (12.5.19)，得到：

$$\hat{\mu}=\frac{\sum_{i=1}^{r}\hat{Z}_i\bar{X}_i}{\sum_{i=1}^{r}\hat{Z}_i}=\frac{0.84\times223.08+0.93\times200}{0.84+0.93}=210.95$$

于是两个被保险团体的个体平均信度保费估计值分别为：

$$\hat{Z}_1\bar{X}_1+(1-\hat{Z}_1)\hat{\mu}=0.84\times223.08+(1-0.84)\times210.95=221.18$$

$$\hat{Z}_2\bar{X}_2+(1-\hat{Z}_2)\hat{\mu}=0.93\times200+(1-0.93)\times206.31=200.72$$

于是，两个被保险团体下一年要交的总的信度保费估计分别为 $221.18\times70=15\ 482.6$ 和 $200.72\times100=20\ 072$。

（3） 此时，$X_{13}=15\ 000/80=187.5$，$\bar{X}_1=(11\ 000+15\ 000)/130=200$，其他不变，经计算可得 $\hat{\mu}=200$，$\hat{v}=16\ 888.57$，

$$\hat{a}=\left[\sum_{i=1}^{r}m_i(\bar{X}_i-\bar{X})^2-\hat{v}(r-1)\right]/\left(m-m^{-1}\sum_{i=1}^{r}m_i^2\right)$$

$$=\frac{130\times(200-200)^2+345\times(200-200)^2-16\ 888.57\times1}{475-(130^2+345^2)/475}$$

$$=-89.43$$

此时 $\hat{a}<0$，因此取 $\hat{Z}_1=\hat{Z}_2=0$，两个团体中的个体平均保费估计均等于 $\hat{\mu}=200$。两个被保险团体总的信度保费估计值分别为 14 000 和 200 000。 ■

12.5.2 半参数估计

在某些情况下，我们不能准确的得到风险参数 Θ 的参数分布形式，但是可以知道当给定 $\Theta=\theta$ 时 X 的条件分布的参数形式，并且有理由确认此条件分布是合理的。这时可应用半参数估计方法。首先根据 $X|\Theta$ 的分布函数得到 $\mu(\Theta)=E[X|\Theta]$ 与 $v(\Theta)=Var[X|\Theta]$ 之间的关系，然后利用 $Var[X]=v+a$ 求出结构参数 μ，v 和 a 的估计值，最后求出信度估计。下面看一个例子。

【例 12－20】 假设某年机动车辆险索赔次数分布如表 12－4 所示。

表 12－4

索赔数	0	1	2	3	4	合计
被保险人数	1 245	211	25	5	3	1 489

基于这个数据计算每个被保险人下一年索赔次数的信度估计。假设索赔次数服从条件泊松分布。已知 X 代表一个被保险人一年内的索赔次数，X 服从均值为 Θ 的泊松分布。经验数据符合 Bühlmann 模型的模式，即对于 r 个被保险人，每人都有 n 年的观察数据，且每年只观察一次，同一个被保险人的数据关于风险参数都是条件独立同分布的。第 i 个被保险人在第 j 年的索赔次数为 X_{ij}。现在要估计第 i 个被保险人在下一年的信度估计 $\hat{X}_{i,n+1}$。

解： 这是一个 Bühlmann 模型，有 1 489 个被保险人，每人有 $n_i=1$，风

险量 $m_{ij}=1$。对第 i 个投保人（$i=1,\cdots,1\ 489$）假设 $X_{i1}\mid\Theta_i=\theta_i$ 服从均值为 θ_i 的泊松分布。据泊松分布的性质，有

$$\Theta=E[X\mid\Theta]=\mu(\Theta)=Var[X\mid\Theta]=v(\Theta)$$

也就是说，

$$\mu=E[\mu(\Theta)]=E[v(\Theta)]=v$$

类似于非参数情形，采用

$$\bar{X}=\frac{1}{1\ 489}(\sum_{i=1}^{1\ 489}X_{i1})=\frac{1}{1\ 489}(0\times 1\ 245+1\times 211+3\times 25+4\times 3)=0.1934$$

并且对于任何分布，均有

$$Var[X_{i1}]=E[Var(X_{i1}\mid\Theta)]+Var[E(X_{i1}\mid\Theta)]=v+a=\mu+a$$

因此 $v+a$ 的一个无偏估计是样本方差

$$\frac{\sum_{i=1}^{1\ 489}(X_{i1}-\bar{X})^2}{1\ 488}=\frac{1}{1\ 488}[1\ 245\times(0-0.1934)^2+\cdots+2\times(4-0.1934)^2]$$
$$=0.234$$

故有 $\hat{a}=0.234-0.1934=0.040$ 和 $\hat{k}=0.1934/0.040=4.76$，信度因子 $\hat{Z}=1/(1+4.76)=0.17$。于是被保险人索赔次数的信度估计值是 $0.17X_{i1}+0.83\ (0.193)$ 其中 X_{i1} 根据不同投保人取值为 0，1，2，3，4。■

12.5.3　参数估计

当我们既可以假设给定 $\Theta=\theta$ 时 X 的条件分布的参数形式，并且有理由确认假设合理时，应该使用参数估计方法。这时，我们可以采用统计上的方法进行处理，极大似然估计是最常用的方法。

对于经验数据 $\vec{X}_i=(X_{i1},X_{i2},\cdots,X_{in_i})$ 和风险参数 Θ_i，假设 $X_{i1}\mid\Theta_i$，$X_{i2}\mid\Theta_i,\cdots,X_{in_i}\mid\Theta_i$ 独立同分布，并且已知其密度函数 $f_{X_{ij}\mid\Theta_i}(x_{ij}\mid\theta_i)$，$i=1,2,\cdots,r$，$j=1,2,\cdots,n_i$，以及 $\Theta_1,\Theta_2,\cdots,\Theta_r$ 独立同分布，密度函数为 $\pi(\theta_i)$，$i=1,2,\cdots,r$。我们可以写出 $\vec{X}_i=(X_{i1},X_{i2},\cdots,X_{in_i})$ 的联合密度函数：

$$f_{\vec{X}_i}(\vec{x}_i)=\int_{\Theta_i}[\prod_{j=1}^{n_i}f_{X_{ij}\mid\Theta_i}(x_{ij}\mid\theta_i)]\pi(\theta_i)d\theta_i \tag{12.5.20}$$

由于不同被保险人之间的损失是独立的，我们可以写出似然函数：

$$L=\prod_{i=1}^{r}f_{\vec{X}_i}(\vec{x}_i)$$

以及对数似然函数：

$$l=\ln L=\sum_{i=1}^{r}\ln f_{\vec{X}_i}(\vec{x}_i)$$

极大似然方法就是通过最大化 L 或 l 得到结构参数的估计值。下面通过一个例子进行介绍。

【例 12-21】 已知经验数据符合 Bühlmann 模型的模式，即 $n_1 = n_2 = \cdots n_r = n$，且 $m_{ij} = 1$，$i = 1, 2, \cdots, r$，$j = 1, 2, \cdots, n$。并且假设 $X_{ij} | \Theta_i \sim N(\Theta_i, v)$，以及 $\Theta_i \sim N(\mu, a)$，即

$$f_{X_{ij}|\Theta_i}(x_{ij} | \theta_i) = (2\pi v)^{-1/2} \times \exp\left[-\frac{1}{2v}(x_{ij} - \theta_i)^2\right], \quad -\infty < x_{ij} < \infty$$

$$\pi(\theta_i) = (2\pi a)^{-1/2} \times \exp\left[-\frac{1}{2a}(\theta_i - \mu)^2\right], \quad -\infty < x_{ij} < \infty$$

求参数 μ，v 和 a 的极大似然估计。

解： 根据式（12.5.20），求出

$$\begin{aligned} f_{\tilde{X}_i}(\tilde{x}_i) &= \int_{\Theta_i} \left[\prod_{j=1}^{n} f_{X_{ij}|\Theta_i}(x_{ij} | \theta_i)\right] \pi(\theta_i) d\theta_i \\ &= \int_{\Theta_i} \left\{\prod_{j=1}^{n} \left\{(2\pi v)^{-1/2} \times \exp\left[-\frac{1}{2v}(x_{ij} - \theta_i)^2\right]\right\}\right\} \times (2\pi a)^{-1/2} \\ &\quad \times \exp\left[-\frac{1}{2a}(\theta_i - \mu)^2\right] d\theta_i \end{aligned}$$

化简、并忽略那些不含 μ，v 和 a 的项：

$$f_{\tilde{X}_i}(\tilde{x}_i) \propto v^{-n/2} a^{-1/2} \int_{\Theta_i} \exp\left[-\frac{1}{2v}\sum_{j=1}^{n}(x_{ij} - \theta_i)^2 - \frac{1}{2a}(\theta_i - \mu)^2\right] d\theta_i \tag{12.5.21}$$

此时若直接计算，运算量较大，我们考虑 $\bar{X}_i$ 的密度函数，根据式（12.5.20），可以求出

$$\begin{aligned} f(\bar{x}_i) &= \int_{\Theta_i} f_{\bar{X}_i|\Theta_i}(\bar{x}_i | \theta_i) \pi(\theta_i) d\theta_i \\ &= \int_{\Theta_i} (4\pi^2 av/n)^{-1/2} \exp\left[-\frac{n}{2v}(\bar{x}_i - \theta_i)^2 - \frac{1}{2a}(\theta_i - \mu)^2\right] d\theta_i \end{aligned}$$

现在比较 $f(\tilde{x}_i)$ 与 $f(\bar{x}_i)$，由于

$$\sum_{j=1}^{n}(x_{ij} - \theta_i)^2 = \sum_{j=1}^{n}(x_{ij} - \bar{x}_i + \bar{x}_i - \theta_i)^2 = \sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2 + n(\bar{x}_i - \theta_i)^2$$

所以，

$$f_{\tilde{X}_i}(\tilde{x}_i) \propto v^{-(n-1)/2} \times \exp\left[-\frac{1}{2v}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2\right] f(\bar{x}_i) \tag{12.5.22}$$

进而，

$$L = \prod_{i=1}^{r} f_{\tilde{X}_i}(\tilde{x}_i) \propto v^{-r(n-1)/2} \times \exp\left[-\frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2\right] \prod_{i=1}^{r} f(\bar{x}_i)$$

接下来求 $f(\bar{x}_i)$。由于 $X_{ij} | \Theta_i \sim N(\Theta_i, v)$，可得 $\bar{X}_i | \Theta_i \sim N(\Theta_i, v/n)$；又由于 $\Theta_i \sim N(\mu, a)$，可以求出 $\bar{X}_i \sim N(\mu, a + v/n)$，即

$$f(\bar{x}_i) = (2\pi\omega)^{-1/2} \exp\left[-\frac{1}{2\omega}(\bar{x}_i - \mu)^2\right]，\text{其中 } \omega = a + v/n$$

也就是说，

$$L \propto v^{-r(n-1)/2}\omega^{-r/2} \times \exp\left[-\frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 - \frac{1}{2\omega}\sum_{i=1}^{r}(\bar{x}_i-\mu)^2\right]$$

现在估计 μ，v 和 ω：

$$L \propto L_1(v)\ L_2(\mu,\omega)，\text{即 } l = l_1 + l_2$$

其中，

$$L_1(v) = v^{-r(n-1)/2} \times \exp\left[-\frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right]$$

$$L_2(\mu,\omega) = \omega^{-r/2} \times \exp\left[-\frac{1}{2\omega}\sum_{i=1}^{r}(\bar{x}_i-\mu)^2\right]$$

$$l_1 = \ln L_1(v) = \frac{-r(n-1)}{2}\ln v - \frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2$$

$$l_2 = \ln L_2(\mu,\omega) = -\frac{r}{2}\ln\omega - \frac{1}{2\omega}\sum_{i=1}^{r}(\bar{x}_i-\mu)^2$$

对 l 求偏导，得：

$$\begin{cases}\dfrac{\partial l}{\partial\mu} = \dfrac{1}{\omega}\displaystyle\sum_{i=1}^{r}(\bar{x}_i-\mu) \\ \dfrac{\partial l}{\partial v} = \dfrac{-r(n-1)}{2}\dfrac{1}{v} + \dfrac{1}{2v^2}\displaystyle\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 \\ \dfrac{\partial l}{\partial\omega} = -\dfrac{r}{2}\dfrac{1}{\omega} + \dfrac{1}{2\omega^2}\displaystyle\sum_{i=1}^{r}(\bar{x}_i-\mu)^2\end{cases}$$

为最大化 l，令上述偏导为零，得到估计值：

$$\hat{\mu} = \frac{\sum_{i=1}^{r}\bar{x}_i}{r} = \bar{X},\quad \hat{v} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2}{r(n-1)},\quad \hat{\omega} = \frac{\sum_{i=1}^{r}(\bar{x}_i-\bar{X})^2}{r}$$

根据 $\omega = a + v/n$ 求出：

$$\hat{a} = \hat{\omega} - \hat{v}/n = \frac{\sum_{i=1}^{r}(\bar{x}_i-\bar{X})^2}{r} - \frac{\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2}{rn(n-1)}$$

这几个估计值似曾相识，本例中得到的 $\hat{\mu}$ 和 $\hat{v}$ 与在 Bühlmann 模型的非参数估计中得到的估计值一致，而 $\hat{a}$ 的差异也只在第一项的分母。

习　题

1. 完全可信条件要求 $\bar{X}$ 在 $0.05E(\bar{X})$ 范围内波动的概率为 0.9，现在有新的标准，要求 $\bar{X}$ 在 $kE(\bar{X})$ 范围内波动的概率为 0.95 。求出 k 的值使这两种标准得到的风险数不变。

2. 基于样本数 $n=100$ 的部分可信因子 $z=0.4$，至少需要增加多少样本

数使 z 增加到 0.5？

3. 每一时期的总理赔额 S 服从复合泊松分布，理赔强度的密度函数为 $f(y)=5y^{-6}$，$y>1$。样本数的完全可信标准要求 S 在 $0.05E(S)$ 范围内波动的概率为 0.9 。如果相同的风险数运用的频数变量 N，则每一个风险期的理赔次数在 $100r\%E(N)$ 内波动的概率为 0.95，试确定 r 的值。

4. 理赔次数的概率分布函数为 $p(x)=\binom{m}{x}q^x(1-q)^{m-x}$，$x=0, 1,\cdots, m$，理赔次数在 $0.01E(X)$ 范围内波动的概率为 0.95，其中完全可信条件下，$E(X)$ 为 34 574，求 q。

5. 设 X 表示理赔额，服从均值为 5 的指数分布，假定 $r=0.05$，$p=0.9$，求理赔额期望估计值 $\bar{X}$ 的完全可信条件。

6. 假设个体风险的索赔次数服从泊松分布，每次索赔额的变异系数为 2，$\alpha=0.1$，$r=0.05$，当个体风险的经验总索赔次数为多少时，用样本赔付额数据估计索赔强度的可信度为 100%？

7. 假设一年内的理赔次数服从均值为 θ 的泊松分布，其先验密度为 $\pi(\theta)=\frac{e^{-\theta}}{1-e^{-k}}$，$0<\theta<k$，每年零索赔的非条件概率为 0.575，试确定 k 的值。

8. 在一个保单组合中，每一个被保险人每年最多只发生一次理赔，其发生概率为 q，先验密度为 $\pi(q)=\frac{q^3}{0.07}$，$0.6<q<0.8$，一个随机抽取的被保险人在第一年理赔一次，在第二年无理赔，对于该被保险人，试确定其后验概率。

9. 对于某一特定风险，一年之内的理赔次数服从均值为 p 的伯努利分布，p 的先验概率分布为 [0, 1] 上的均匀分布，计算得到的贝叶斯信度估计值是观察理赔额的 1/5 时，则理赔额为 0 的年数是多少？

10. 一个保单组合有 100 个独立的个体，其中 25 个个体的理赔限额为 5 000，25个的理赔限额为10 000，50个的理赔限额为 20 000。在分类以前，这些风险个体拥有相同的理赔额分布，即服从参数分别是 $\theta=5\ 000$，$\alpha=2$ 的帕累托分布，在分类以后，根据理赔报告可以显示出每一个范围的风险数，但是区分不开每一次理赔的理赔限额。这个报告准确显示了一个随机选择的理赔，位于 9 000 ~ 11 000 范围内，试确定该个体属于理赔限额为 10 000 的概率。

11. 两个盒子每个里面都装了形状相同的 10 个球。第一个盒子里面有 5 个红球和 5 个白球，第二个盒子里面有 2 个红球和 8 个白球，每个球被抽中的概率是相等的。现随机抽选一个盒子，两个盒子等概率被抽中；从这个盒子中随机选出一个球，放回原盒子后再从该盒子中随机选出一个球。假设第

一个被抽中的球是红色的，那么第二个被抽中的球也是红色的概率是多少？

12. 在观察到任何理赔以前，你认为理赔额的大小服从参数为 $\theta=10$，$\alpha=1$，2 或者 3 的帕累托分布，三种情况等概率。现在观察到一个随机抽取的样本理赔额为 20 ，试确定该样本点下次理赔额大于 30 的后验概率。

13. 一个完全独立个体的风险集可分为两类，每一类拥有相同的样本数。在类别 1 中，每一年的理赔数服从均值为 5 的泊松分布；在类别 2 中，每一年的理赔数服从参数为 $m=8$，$q=0.55$ 的二项分布。一个随机选择的风险个体在第一年有 3 次理赔，在第二年有 r 次理赔，在第三年有 4 次理赔。Buhlmann 信度估计在第四年的理赔数为 4.6019。求 r。

14. 某保险公司售出一个保单组合，过去的经验显示平均的理赔频率为 0.425，期望值的方差为 0.37，方差的均值为 1.793。现在从保单组合中随机选择一种被保险人，该种类别的被保险人中再选出 9 个个体，一共有 7 次理赔。现在从这种类别中再选出 5 个个体，求出这 5 个个体总理赔数的 Buhlmann 信度估计。

15. 已知两个风险 A 和 B 的损失金额服从表 12－5 所示的分布。

其中风险 A 发生损失的概率是风险 B 的两倍。如果已知某个风险在某次事故中的损失额为 300，求该风险下次损失额的 Buhlmann 信度估计。

表 12－5

损失额	风险 A 的概率分布	风险 B 的概率分布
300	0.5	0.6
3 000	0.3	0.3
70 000	0.2	0.1

16. 考虑一个由团体保单形成的保单组合。对整个保单组合而言，平均每个被保险人的期望纯保费为 2 400。对于不同的团体保单，平均每个被保险人的纯保费是不同的，不同假设均值之间的方差为 500 000 。对于同一个团体保单，不同被保险人的纯保费也存在差异（用组内方差表示），所有团体保单的过程方差的均值为 250 000 000。假设一份团体保单上年的索赔经验如下：被保险人数为 240 人，平均每个被保险人的经验纯保费为 3 000。计算该团体保单下每个被保险人的信度纯保费。

17. 假设风险集合中只有两个规模相等的个体风险，对每个风险的观察期均为 3 年，第一个风险的经验损失为：3，5，7；第二个风险的经验损失为：6，12，9 。计算这两个风险的 Bühlmann 信度保费。

18. 理赔次数服从均值为 m 的泊松分布，理赔额服从均值为 $20m$ 方差为 $400m^2$。m 的密度函数为 $f(m)=\dfrac{m^2e^{-m}}{2}$，$0<m<\infty$，其中对于任何 m，理赔额和理赔次数的分布是独立的。确定总理配额组内方差的期望。

19. 在大量的商业被保险人中你得到了如下数据：每个被保险人的损

失是独立的，并且拥有相同的均值和方差，均值为25，假设期望的方差为50，条件方差的期望为10 000。现随机选择一个被保险人得到表12－6所列经验数据。试确定每个被保险人的Bühlmann－Straub保费。

表12－6

年度	损失的均值	被保险人数
1	20	1 000
2	15	750
3	10	600

20. 一年度两类风险的累计损失分布图表12－7所示。随机选择一个风险个案，观察到在前两个年度的损失都为0。确定该个案在第三年度的贝叶斯信度估计值。

表12－7

风险	累计损失		
	0	50	1 000
A	0.8	0.16	0.04
B	0.6	0.24	0.16

21. 在第20题的条件下，确定该个案在第三年度的Bühlmann信度保费。

22. 一个保险公司有两组投保单。在头四个保单年度总理赔额如表12－8所示（单位为百万元）。假设这两组保单有相同数目的被保险人，根据Bühlmann模型计算每一组在第五年的经验贝叶斯信度保费。

表12－8

累计总保费	保单年度			
组别	1	2	3	4
1	4	10	8	6
2	12	14	13	13

23. 一个保险公司有两组保单，头三年的累计理赔额如表12－9所示，假设这两组拥有相同的被保险人数。求第一组保单在第四年的Bühlmann信度保费。

表 12－9

累计总理赔	保单年度		
组别	1	2	3
1	5	8	11
2	11	13	12

24. 在第 23 题的条件下，利用个体样本的信度加权平均计算 μ，找出第一组保单在第四年的 Bühlmann 信度保费。

25. 在第 23 题的条件下假设第二组保单在保单年度 1、2、3 的累积总理赔为 2、8、14，计算第一组保单在第四年的 Bühlmann 信度保费。

第十三章 随机模拟

学习目标

- □ 了解均匀分布随机数产生的原理
- □ 熟悉一般分布和正态随机向量的随机数产生方法，掌握泊松分布、正态分布随机数产生方法
- □ 了解 Bootstrap 方法的基本原理，熟悉使用 Bootstrap 方法计算均方误差
- □ 了解 MCMC 模拟的基本原理
- □ 熟练运用随机模拟方法解决精算中的实际问题

§13.1 引 言

客观世界的某些现象之间存在着某种相似性，因而可以从一种现象出发研究另一种现象。比如在分析一个系统时，可先构造一个与该系统相似的模型，通过在模型上进行试验来研究原系统，这就是模拟（Simulation）。随机系统可以用概率模型来描述并进行试验，称为**随机模拟方法**，又称蒙特卡罗（Monte Carlo）方法或统计试验法。

随机模拟在保险精算中的用途非常广泛，既可用于确定性问题，又可用于随机问题的处理。当某一问题用传统的方法处理有较大难度或计算过于繁杂时，就可以采用随机模拟方法。例如在分析保险公司的资产与负债配比策略、聚合理赔风险等问题时，都可用到随机模拟方法。一般地说，在以下几种情况下，随机模拟方法将发挥其独特作用：

1. 无论从费用还是从时间上均难以对风险系统进行大量实测；
2. 由于实际风险系统的损失后果严重而不能进行实测；
3. 难以对复杂的风险系统构造精确的解析模型；
4. 用解析模型不易求解；
5. 对解析模型进行验证。

随机模拟的基本步骤是：（1）建立恰当模型；（2）设计试验方法；（3）从一个或多个概率分布中重复生成随机数；（4）分析模拟结果。在这四个基本模拟步骤中，（1）、（2）和（4）与所研究的特定问题相关性很强，而（3）随机数的生成则是任何模拟的基本要素。因此，本章将讨论几种常用分布的随机数生成方法，同时对模拟样本的容量问题进行初步的分

析，阐述 Bootstrap 模拟和 MCMC 模拟方法及应用，最后通过几个简单例子来说明模拟方法在保险精算中的应用。

实际问题中，由于影响系统的因素很多，在建立系统模型时常作某些假设，而对某些因素则需要忽略不计或将几个因素合并考虑。为了使分析结果更加切合实际，还要进行灵敏度分析，即在固定其他因素不变的条件下分别对感兴趣的因素作变动，并观察其对最后结果的影响。这样就能对真实结果与分析结果的偏离有所认知并预先采取相应的对策。

§13.2　均匀分布随机数与伪随机数

13.2.1　原　理

产生均匀分布的随机数是进行随机模拟的基础。为了清楚地说明这一点，以计算定积分 $\int_0^1 f(x)dx$ 为例，由微积分学的基本原理可知，若被积函数 $f(x)$ 的原函数 $F(x)$ 存在，则可以采用牛顿—莱布尼兹公式求其值。但实际上被积函数往往比较复杂难以求得其原函数，尤其是多变量情形下的重积分计算问题。这时，也可以从另一个不同的角度来考虑这个问题。

设 U 是[0，1]区间上均匀分布的随机变量，利用被积函数 $f(x)$ 构造一个新的随机变量 $\xi=f(U)$，由概率论的基本原理可知：$E(\xi)=\int_0^1 f(x)dx$，所以 $\int_0^1 f(x)dx$ 的近似计算问题转变为数学期望 $E(\xi)$ 的估计问题。

由数理统计学的基本原理可知，随机变量 ξ 的样本均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 是对其数学期望 $E(\xi)$ 的较优的估计，但是 ξ 本身无法直接观测，我们可以通过 U 的观测样本来构造 ξ 的样本。具体步骤如下：

利用某种手段产生[0，1]上相互独立且服从均匀分布的 n 个随机数 u_1，u_2，…，u_n，即为 U 的样本，则所求定积分 $\int_0^1 f(x)dx$ 的近似计算公式为：

$$\int_0^1 f(x)dx=\frac{1}{n}\sum_{i=1}^{n}f(u_i) \tag{13.2.1}$$

该近似方法对于多重积分的计算特别适用。这也是随机模拟基本思想的一种体现。

在上面的方法中如何产生[0，1]上均匀分布的 n 个随机数 u_1，u_2，…，u_n 是随机模拟的关键。现已有很多方法用于产生均匀分布的随机数，对这些方法通常要求满足以下条件：

1. 统计特性好：要求随机数的分布具有一定的均匀性，即所得数列的

统计性质与[0，1]上均匀分布的样本相同，或至少相当近似。

2. 循环周期长：一般的随机模拟都要有成千上万个随机数，而任何随机数生成方法都会出现周期性的重复，也称为**随机数的循环周期**。若随机数在较短的周期内就发生循环，即使增加模拟的次数也无法解决这个问题。

3. 计算简便：需要相对较少的步骤产生一个随机数，以使计算本身比较方便。

13.2.2 产生均匀分布随机数的几种方法

产生均匀分布随机数的方法，大致可分为三类。

第一类方法为检表法，就是事先将一些服从均匀分布的随机数编成表格，需要时直接从这一表格中取用。当进行模拟时，需先将该表输入计算机。这样做要占用一定的存储单元，现已较少采用。只是在进行手工模拟时，仍可使用该方法。需要注意的是使用随机数表时必须保持读表取数的随机性。

第二类方法是物理方法，就是把具有随机性质的物理过程变换为随机数，比如以放射性物质为随机源的放射随机数发生器等等，可以获得真正的随机数。将发生器与计算机连接，模拟时随时调用。但由于增加了硬件设备，因而增加了费用和维护工作，又无法对模拟问题进行复算检查，故降低了其使用价值。

第三类方法是数学方法，是被广泛采用的一种方法，实质上是利用了计算机的算术和逻辑运算的能力。首先给定一组初值：u_0，u_{-1}，…，u_{-k}，$u_{-(k+1)}$，然后用一个适当的数字递推式：

$$u_n = g(u_{n-1}, u_{n-2}, \cdots, u_{n-k}), \qquad n = 1, 2, \cdots \tag{13.2.2}$$

可逐步求出 u_1，u_2，…，从而产生具有均匀总体随机子样性质的随机数。

由于计算机所能存贮的字节数有限，只能表示有限个不同的数，不能产生真正连续分布的随机数，而且得到的序列是用算法产生的，这些生成的数在本质上是确定的，因此，在采用式（13.2.2）递推产生的序列达到一定长度后，或退化为零，或周期性循环出现。所以这样的随机数只能称为**伪随机数**。

产生伪随机数的方法很多，下面介绍最常用的算法**乘同余法**。

给定的初值 w_0（又称**种子**），w_0 为自然数，取正整数 k_1 和 m，按下面的公式递推求出：

$$\begin{aligned} w_n &= k_1 w_{n-1} (\bmod m) \\ u_n &= w_n / m \end{aligned} \qquad (0 \leqslant w_n \leqslant m) \tag{13.2.3}$$

其中 mod 表示模数运算。显然，有 $0 \leqslant u_n < 1$ 成立。

乘同余法又称为**一阶线性同余法**，它是**混合同余法**的一个特例。用混

合同余法产生随机数的递推式为：

$$\begin{aligned} w_n &= k_1 w_{n-1} + k_2 w_{n-2} + \cdots + k_s w_{n-s} + I(\mathrm{mod}\, m) \\ u_n &= w_n / m \end{aligned} \quad (0 \leqslant w_n \leqslant m) \quad (13.2.4)$$

这里的 k_1、m 和初值 w_0 都是正整数。

从上面的定义容易看出，用混合同余法（包括乘同余法）产生随机数时，由于 $0 \leqslant w_n \leqslant m$，所以 w_n 是有限的，故随机数一定存在某个正的周期，也就是说从某个正整数 p 开始有：

$$w_{n+p} = w_n,\ (n = 1, 2, \cdots) \quad (13.2.5)$$

因此，从 $p+1$ 项开始，由混合同余法产生的数列开始重复，产生周期现象。周期越长，不重复的随机数也越多，也越符合均匀总体随机子样的统计性质。J. G. Skellam 建议在乘同余法中取 $m = 999\ 563$ 及 $k_1 = 470\ 001$。

也可以考虑高阶的同余法，二阶、三阶线性同余法的递推公式如下：

$$\begin{aligned} w_n &= k_1 w_{n-1} + k_2 w_{n-2} + I(\mathrm{mod}\, m) \\ u_n &= w_n / m \end{aligned} \quad (13.2.6)$$

$$\begin{aligned} w_n &= k_1 w_{n-1} + k_2 w_{n-2} + k_3 w_{n-3} + I(\mathrm{mod}\, m) \\ u_n &= w_n / m \end{aligned} \quad (13.2.7)$$

表 13－1 中列出了 Skellam 建议的 m、k_1、k_2 和 k_3 的值，阶数为 s 的同余法对应的周期为 $m^s - 1$。

表 13－1　高阶同余法的参数选择示范

s	m	k_1	k_2	k_3
1	999 563	470 001		
2	998 917	366 528	508 531	
2	999 563	254 754	529 562	
3	997 783	360 137	519 815	616 087
3	997 783	286 588	434 446	388 251

上述各种产生随机数的数学方法都是基于某个递推公式，故在这些随机数之间必然存在一定的相关关系，因此在使用这些随机数之前必须对其进行统计检验。对[0，1]区间均匀分布的随机数，可采用卡方检验法检验其均匀性，用相关系数法检验其独立性。如果经统计检验随机数不符合要求，就需要改变生成方法中的参数或对随机数进行改进。如果所得数列能通过统计检验，则认为其统计特性较好。研究表明，用乘同余法所得的随机数列，性质较为良好。

有了[0，1]区间均匀分布的随机数，可以自然地导出任意区间[a，b]

上均匀分布的随机数 s_n，只要将[0，1]区间均匀分布随机数 u_n 经过线性变换 $s_n=a+(b-a)u_n$ 即可。因此，产生[0，1]区间上均匀分布的随机数是随机模拟问题的基础。也正是由于这个原因，在各种应用统计软件中都预设了产生[0，1]区间上均匀分布的随机数的程序，可以直接调用。如最常用的表格计算工具 Microsoft Excel 中，只要启动“函数”功能中的“RAND”函数，就能获得[0，1]区间上均匀分布的随机数。当然，在计算机上使用函数产生的随机数也是伪随机数。

§13.3 一般分布随机数

13.3.1 几种主要方法的介绍

在均匀分布随机数的基础上，就可以对服从各种概率分布的随机变量进行模拟，进而模拟各种复杂的实际系统。在均匀分布随机数基础之上产生其他分布随机数的数学问题可表达为：已有来自均匀分布总体 $U[0,1]$ 的随机数 u，对于随机变量 $X\sim F(x)$，试求来自于总体 X 的子样 x。

常见的方法有四种：反函数法、取舍法、Box - Muller 方法和极方法，它们可分别用于产生各种分布的随机数。现对它们分别介绍如下：

1. *反函数法*。设随机变量 X 的分布函数为 $F(x)$，定义

$$F^{-1}(u)=\inf\{x:F(x)\geqslant u\},\qquad 0<u<1 \tag{13.3.1}$$

假设随机变量 U 服从(0，1)上的均匀分布，则随机变量 $F^{-1}(U)$ 的分布函数是 $F(x)$，即在分布相等的意义上有 $X=F^{-1}(U)$。

要证明 x 为所要求的随机数，只要证明所构造的随机变量 $F^{-1}(U)$ 具有分布函数 $F(x)$ 即可，事实上，$X=F^{-1}(U)$ 的分布函数为：

$$\begin{aligned}P(X\leqslant x)&=P(F^{-1}(U)\leqslant x)=P(F(F^{-1}(U))\leqslant F(x))\\&=P(U\leqslant F(x))=F(x)\end{aligned}$$

具体步骤如下：

(1) 由 $U(0,1)$ 抽取 u；

(2) 计算 $x=F^{-1}(u)$，其中 F^{-1} 如式（13.3.1）中定义。

以上步骤的含义是，若已知分布函数 $F(x)$，只要产生一个均匀分布随机数 u，对应的取分布 $F(x)$ 的 u 分位点即可。形象地说就是，由均匀分布随机数 u 确定了纵坐标，然后根据分布 F，找对应的原象（横坐标）。如果纵坐标 u 有多个原象（多个横坐标的函数值相等），由式（13.3.1），则取数值最小的原象 x_0（横坐标），见图 13 -1。如果纵坐标 u 没有对应的原象（u 不在分布函数的值域中），则取值域中大于 u 且最接近 u 的纵坐标 v 的原象 x_1，即跳跃发生时的横坐标，见图 13 -2。

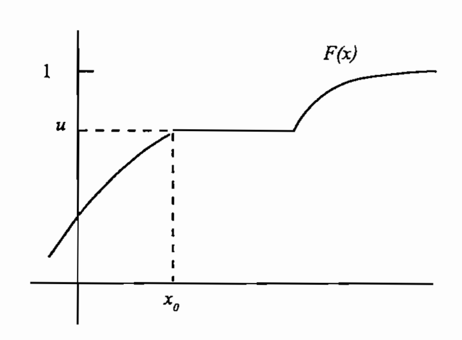


图 13－1　一个特殊情况时分位点的取法

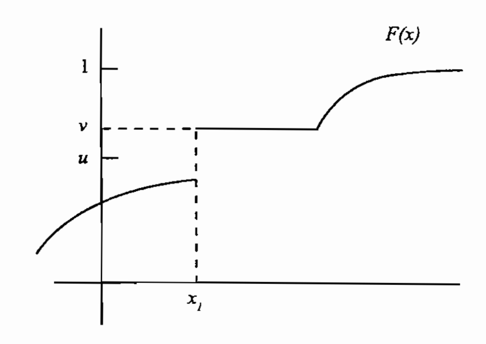


图 13－2　有跳跃时分位点的取法

如果 $F(x)$ 是连续且严格单调的，则 $F^{-1}(u)$ 为 F 的反函数。当分布形式简单时，可以直接计算反函数，如指数分布。但有些分布的反函数没有显式，如正态分布，这时只能数值计算或采用其他方法。

2. 取舍法。若 X 的分布密度函数 $f(x)$ 可以表示为 $f(x)=c\cdot h(x)\cdot g(x)$，其中 $c\geqslant 1$ 为常数，$h(x)$ 是一个相对简单（易于获得分布函数逆函数的解析表达式）的密度函数，因此假定已有现成的方法获得分布密度为 $h(x)$ 的随机数，另外有 $0<g(x)\leqslant 1$。首先产生均匀分布随机数 u 和密度函数为 $h(x)$ 的随机数 y，若 $u\leqslant g(y)$，则令 $x=y$，否则重新生成均匀分布的随机数 u 和 $h(x)$ 的随机数 y，直至满足条件。

为了证明 x 是所要求的随机数，只需证明 X 的分布密度为 $f(x)$。考虑分布函数为 $F(x)$ 的 X、密度函数为 $h(x)$ 的 Y 以及均匀分布的 $U[0,1]$，三者的关系式如下：

$$F(x)=P(X\leqslant x)=P(Y\leqslant x\mid U\leqslant g(Y))$$
$$=\frac{P(Y\leqslant x,\ U\leqslant g(Y))}{P(U\leqslant g(Y))}$$

由于 $P(U\leqslant g(Y))=\int_{-\infty}^{\infty}P(U\leqslant g(s)\mid Y=s)h(s)ds=\int_{-\infty}^{\infty}g(s)h(s)ds=c^{-1}$

以及 $P(Y\leqslant x,\ U\leqslant g(Y))=\int_{-\infty}^{\infty}P(Y\leqslant x,\ U\leqslant g(s)\mid Y=s)h(s)ds$

$$=\int_{-\infty}^{x}P(U\leqslant g(s))h(s)ds$$
$$=\int_{-\infty}^{x}g(s)h(s)ds$$

因此有 $F(x)=c\int_{-\infty}^{x}g(s)h(s)ds$，亦即 X 的密度为 $f(x)$。

3. Box-Muller 方法。首先产生[0，1]区间上两个独立的均匀分布随机数 u_1 和 u_2，令：

$$x_1 = (-2\ln u_1)^{\frac{1}{2}}\cos(2\pi u_2)$$
$$x_2 = (-2\ln u_1)^{\frac{1}{2}}\sin(2\pi u_2) \tag{13.3.2}$$

则 x_1 和 x_2 是两个独立的、服从标准正态分布的随机数。

Box-Muller 方法的原理是，由式（13.3.2）可以推出以下关系式：

$$u_1 = e^{-\frac{1}{2}(x_1^2+x_2^2)},\ u_2 = \frac{1}{2\pi}\arctan\left(\frac{x_2}{x_1}\right)$$

则
$$\frac{\partial u_1}{\partial x_1} = -x_1 e^{-\frac{1}{2}(x_1^2+x_2^2)},\ \frac{\partial u_1}{\partial x_2} = -x_2 e^{-\frac{1}{2}(x_1^2+x_2^2)}$$

$$\frac{\partial u_2}{\partial x_1} = \frac{1}{2\pi}\frac{1}{1+\left(\frac{x_2}{x_1}\right)^2}\left(-\frac{x_2}{x_1^2}\right),\ \frac{\partial u_2}{\partial x_2} = \frac{1}{2\pi}\frac{1}{1+\left(\frac{x_2}{x_1}\right)^2}\left(\frac{1}{x_1}\right)$$

这时变换的 Jaccobi 行列式为：

$$\left|\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right| = \left|-\frac{1}{2\pi}e^{-\frac{1}{2}(x_1^2+x_2^2)}\right|$$

由于 U_1，U_2 是[0，1]上独立的均匀分布，可知 X_1，X_2 的联合分布密度为：

$$\frac{1}{2\pi}e^{-\frac{1}{2}(x_1^2+x_2^2)} = \frac{1}{\sqrt{2\pi}}e^{-\frac{x_1^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2}}$$

即 X_1，X_2 是相互独立的 $N(0, 1)$ 随机变量。

4. *极方法*。极方法实际上是 Box-Muller 方法的特例。首先生成在[-1，1]区间上均匀分布的随机数 u_1 和 u_2；令 $v_i = 2u_i - 1$，$i = 1, 2$；以及 $\rho = v_1^2 + v_2^2$。若 $\rho > 1$，重新产生 u_1 和 u_2，否则，令

$$x_1 = v_1\sqrt{\frac{-2\ln\rho}{\rho}},\ x_2 = v_2\sqrt{\frac{-2\ln\rho}{\rho}} \tag{13.3.3}$$

则 x_1，x_2 是相互独立的 $N(0, 1)$ 随机变量的随机数。

若要生成均值为 μ，方差为 σ^2 的正态分布的随机数，可先得到标准正态分布的随机数，然后将该数字乘以 σ 并加上 μ 即可得到 $N(\mu, \sigma^2)$ 随机变量的随机数。

13.3.2 连续随机变量的模拟

【例 13-1】 设随机变量 X 服从指数分布，$f(x) = \frac{1}{\theta}e^{-x/\theta}$，$x > 0$，请利用反函数法，根据[0，1]区间均匀分布的随机数 u_1，u_2，…来表示随机变量 X 的随机数。

解： 指数分布的分布函数为：

$$F(x) = 1 - e^{-x/\theta},\qquad x > 0,\ \theta > 0$$

其反函数为 $F^{-1}(u) = -\theta\ln(1-u)$。因此，要生成均值为 θ 的指数分布的随机数 x，采用以下两个步骤：

第一步：生成[0，1]区间均匀分布的随机数 u；

第二步：令 $x = F^{-1}(u) = -\theta\ln(1-u)$。

值得注意的是，若 U 服从(0，1)均匀分布，则 $1-U$ 仍然是服从(0，1)均匀分布，所以 $1-u$ 仍然是(0，1)均匀分布的随机数。又由于分布函数是单调递增的，小的数值对应的分位点也要小，所以对于较小的均匀分布的随机数 u，如果取 u 对应的分位点，则意味着小的随机数产生小的模拟数；如果取 $1-u$ 对应的分位点，则意味着小的随机数产生大的模拟数。例如在例 13－1 中，可得到另一个指数分布的随机数 $x' = -\theta\ln(1-(1-u)) = -\theta\ln(u)$。

【例 13－2】 利用[0，1]区间均匀分布的随机数 u_1，u_2，…表示服从帕累托分布的随机数。

解： 设 X 服从帕累托分布，则

$$F_X(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha}$$

由 $U = F_X(x)$ 得出 $u_i = 1 - \left(\frac{\theta}{x_i+\theta}\right)^{\alpha}$，所以 $x_i = \theta(1-u_i)^{-\frac{1}{\alpha}} - \theta$ 即为所求 。

【例 13－3】 试产生标准正态分布的随机数。

解： 因为正态分布函数的反函数不容易给出解析表达式，因此不宜采用反函数法。但 Box－Muller 方法和极方法都可以直接应用。

由于正态分布在统计理论中占有重要地位，所以在数理统计理论和实践中常常将标准正态分布的分布函数做成表格，利用标准正态分布表和均匀分布随机数表就可得到标准正态分布的随机数。

具体步骤为：首先得到[0，1]均匀分布的随机数 u，然后在标准正态分布函数表中找到使 $\Phi(x) = u$ 成立的 x 值，则此 x 是标准正态分布函数的一个随机数。比如从均匀分布随机数表中得到随机数 0.02488，查标准正态分布表可知 $\Phi(-1.962) = 0.02488$，即标准正态随机数为 -1.962。这种方法也可在计算机上使用，但存贮标准正态分布表也需要占用一定的空间。

还有一种方法是利用中心极限定理。由中心极限定理，由 n 个[0，1]区间均匀分布的随机变量 $U_1, U_2, \cdots, U_n$ 生成的随机变量：

$$X = \left[\sum_{i=1}^{n} U_i - \frac{n}{2}\right] \Big/ \sqrt{\frac{n}{12}} \qquad (13.3.4)$$

当 n 足够大时近似服从标准正态分布。

所以，首先生成[0，1]区间均匀分布的 n 个随机数 $u_1, u_2, \cdots, u_n$，然后代入式（13.3.4）计算得到的 x 即为标准正态分布的随机数。

若 $n = 12$，则式（13.3.4）简化为：

$$X = \sum_{i=1}^{12} U_i - 6$$

或 $X = U_1 + \cdots + U_6 - (1-U_7) - \cdots - (1-U_{12})$

又由于 U_i 和 $1-U_i$ 都是[0，1]上的均匀分布，也可选取以下的随机数：

$$x = u_1 + \cdots + u_6 - u_7 - \cdots - u_{12} \tag{13.3.5}$$

在对计算速度要求较高，且对正态分布的尾端分布情况要求不高时，这一方法适用。

【例 13－4】 试生成参数为 $\mu=5.0$，$\sigma^2=4.0$ 的对数正态分布的 5 个随机观察值。

解： 若要生成对数正态分布的随机数，可以先生成标准正态分布的随机数，然后经过参数变换生成相关正态分布的随机数，再对结果进行指数变换就可得到要求的对数正态分布随机数。

表 13－2 给出了生成 5 个对数正态随机数的上述各步骤的中间结果。

表 13－2　　　　例 13－4 对数正态分布随机数

随机数 u	标准正态分布随机数 $\Phi^{-1}(u)$	正态分布随机数 $\mu+\sigma\Phi^{-1}(u)=5+2\Phi^{-1}(u)$	对数正态分布随机数 $\exp[5+2\Phi^{-1}(u)]$
0.81525	0.90	6.80	897.8
0.29676	-0.53	3.94	51.4
0.00742	-2.44	0.12	1.1
0.05366	-1.61	1.78	5.9
0.91921	1.40	7.80	2 240.6

【例 13－5】 试生成伽玛分布（参数为 α－正整数，$\gamma>0$）的随机数。

解： 伽玛分布的反函数不能给出解析表达式，这时可根据伽玛随机变量与指数随机变量的关系得到生成伽玛随机数的一种特殊方法。我们知道，α 个参数均为 γ 的指数分布随机变量之和服从伽玛(α，γ)分布，即若 X_i 服从参数为 γ 的指数分布，$1\leqslant i\leqslant\alpha$，则 $\sum_{i=1}^{\alpha} X_i$ 服从伽玛(α，γ)。因此，为了生成伽玛(α，γ)的随机数，可首先生成 α 个均值为 γ 的指数分布随机数：$x_i=-\gamma\ln(1-u_i)$，$1\leqslant i\leqslant\alpha$。这里 u_i 为均匀分布随机数，然后将这 α 个随机数相加得到伽玛(α，γ)的一个随机数 $y_i=\sum_{i=1}^{\alpha} x_i=-\gamma\sum_{i=1}^{\alpha}\ln(1-u_i)$。

【例 13－6】 试生成 $\chi^2(n)$ 分布的随机数。

解： 这里的基本原理是：若 X_1，X_2，…，X_n 是 n 个独立同分布的标准正态分布随机变量，则 $\sum_{i=1}^{n} X_i^2$ 服从自由度为 n 的 χ^2 分布。因此，要生成 $\chi^2(n)$ 分布的随机数，首先生成 n 个标准正态分布随机数，然后将这 n 个随机数的平方求和即为所求的 $\chi^2(n)$ 随机数。

13.3.3　离散随机变量的模拟

设离散型随机变量 X 的概率分布为 $P(X=a_i)=p_i$，$i=1$，…，n。其中

$p_i>0$，$\sum_{i=1}^{n}p_i=1$，并且设 $a_1<a_2<\cdots<a_n$。根据反函数原理，对均匀分布随机数 u，只需找整数 j，满足 $\sum_{i=0}^{j-1}p_i\leqslant u<\sum_{i=0}^{j}p_i$，则 X 的模拟值取 a_j。

有的时候，要将 X 的各种可能事件 a_i 的概率全部进行计算过于繁杂，因此对于不同的分布可采用不同的特定方法。

【例 13－7】 考虑模拟贝努里随机变量的方法。

解： 若 X 为贝努里随机变量，成功概率为 p，失败的概率为 $q=1-p$。通常，若随机数 u 在区间$[p,1]$上，则说明试验失败；否则说明试验成功，即 $u\in[0,p)$。

【例 13－8】 考虑模拟泊松分布的方法。

解： 泊松分布常被用于表示索赔次数的分布，概率分布为：

$$p_k=P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!},\qquad k=0,1,\cdots$$

其中参数 $\lambda=E(X)$表示平均索赔次数。

方法一： 可以直接应用反函数法来生成泊松随机数，记：

$$F_0=e^{-\lambda}$$

$$F_n=\sum_{k=0}^{n}p_k=F_{n-1}+p_n,\qquad n=1,2,\cdots$$

然后执行以下步骤：

（1）生成$[0,1]$区间均匀分布随机数 u；

（2）若 $u<F_0$，则令 $x=0$；

（3）若存在某个 $k>0$ 使得 $F_{k-1}\leqslant u<F_k$，则令 $x=k$。

方法二： 分数乘积法，主要是基于以下事实：若两次相继发生的事件之间的时间间隔系列 T_1，T_2，$\cdots$独立且服从均值为 $1/\lambda$ 的指数分布，则在单位时间段内事件发生的次数 X 服从参数为 λ 的泊松分布（参见第七章）。

记 X 为单位时间段内事件发生的次数，则 X 与 T_i 有以下关系式成立：

$$\sum_{i=0}^{X}T_i\leqslant 1<\sum_{i=0}^{X+1}T_i\quad(T_0=0)\tag{13.3.6}$$

所以首先需要得到 T_i 的随机数，由模拟指数分布的方法知，若 u_i，$i=1$，2，$\cdots$为$[0,1)$区间均匀分布的随机数，则：$t_i=-\frac{1}{\lambda}\ln u_i$ 为 T_i 的随机数。

由式（13.3.6），存在某个 x 满足：

$$\sum_{i=0}^{x}-\frac{1}{\lambda}\ln u_i\leqslant 1<\sum_{i=0}^{x+1}-\frac{1}{\lambda}\ln u_i,\qquad u_0=1$$

两边同乘以 $-\lambda$ 可得：

$$\sum_{i=0}^{x}\ln u_i\geqslant-\lambda>\sum_{i=0}^{x+1}\ln u_i$$

即 $$\prod_{i=0}^{x} u_i \geqslant e^{-\lambda} > \prod_{i=0}^{x+1} u_i \qquad (13.3.7)$$

则满足式（13.3.7）的 x 为泊松分布的一个随机数。

具体的操作方法如下：首先从 0 点开始，若 $e^{-\lambda} > u_1$，则令 $x=0$；否则，继续比较，若 $e^{-\lambda} > u_1 \cdot u_2$，则令 $x=1$；依此继续，直至存在某个 k 首次满足：$e^{-\lambda} > \prod_{i=1}^{k+1} u_i$。所以，这个方法常常需要多个[0，1]均匀分布的随机数。

考虑生成 5 个参数 $\lambda=0.5$ 的泊松分布的随机数。由已知条件有：$e^{-0.5}=0.60653$，首先产生一列[0，1]均匀分布的随机数，例如 0.68379，0.10493，0.81889，0.81953，0.35101，0.16703，0.83946，0.35006，0.20226，…，然后按照上面的方法逐步的选取。计算过程和结果列在表 13－3 中。

表 13－3　　泊松分布的随机数（分数乘积法）

序号	$\prod_{i=0}^{x+1} u_i$	$\prod_{i=0}^{x} u_i$	随机数 x
1	0.68379 × 0.10493 = 0.07175	0.68379	1
2	0.81899 × 0.81953 × 0.35101 = 0.23559	0.81899 × 0.81953 = 0.67119	2
3	0.16703	1	0
4	0.83946 × 0.35006 = 0.29386	0.83946	1
5	0.20226	1	0

这种方法在 λ 较大时，计算过程将变得复杂而且计算量增大，这时也可以考虑采用中心极限定理来减少计算量。

方法三：中心极限定理近似。根据中心极限定理，在 λ 很大时，参数为 λ 泊松分布可用均值为 λ，标准差为 $\sqrt{\lambda}$ 的正态分布来近似。因此可采用如下的方法生成泊松分布随机数：

第一步：生成均匀分布随机数 u；

第二步：计算相应的标准正态随机数 z：

第三步：计算 $\lambda + z\sqrt{\lambda}$；

第四步：按照 4 舍 5 入的原则，选取与 $\lambda + z\sqrt{\lambda}$ 最接近的非负整数，并以此作为泊松分布随机数 x。

例如要生成泊松参数 $\lambda=11$ 的 5 个随机数，计算过程和结果列在表 13－4 中。

表 13-4　泊松分布的随机数（正态近似法）

随机数 u	标准正态随机数 z	$11+z\sqrt{11}$	泊松随机数 x
0.91567	1.38	15.58	16
0.17955	-0.92	7.95	8
0.46503	-0.09	10.70	11
0.92157	1.41	15.68	16
0.14577	-1.06	7.48	7

【例 13-9】 考虑模拟负二项分布的方法（参数为 $k=6$、$p=0.6$）。

解：

方法一：直接采用反函数法方法生成负二项分布的随机数。为此需要首先计算概率分布表，表 13-5 列出了参数为 $k=6$、$p=0.6$ 的负二项分布的概率分布值。然后基于该分布按照以下步骤进行：

第一步：生成[0，1]上均匀分布随机数 u；

第二步：若 $u<0.0467$，取随机数 $x=0$；否则，继续进行，若 $0.0467\leqslant u<0.1587$，取 $x=1$；如此等等。

表 13-5　负二项分布表（$k=6$，$p=0.6$）

j	$P\{N=j\}$	$P\{N\leqslant j\}$
0	0.0467	0.0467
1	0.1120	0.1587
2	0.1568	0.3155
3	0.1672	0.4827
4	0.1505	0.6332
5	0.1204	0.7536
6	0.0883	0.8419
7	0.6050	0.9024
8	0.0394	0.9418
9	0.0245	0.9663
10	0.0147	0.9810
11	0.0085	0.9895
12	0.0048	0.9943
13	0.0027	0.9970
14	0.0015	0.9985
15	0.0008	0.9993
16	0.0004	0.9997
17	0.0002	0.9999
⋮	⋮	⋮

方法二：中心极限定理近似。当负二项分布的参数 k 较大时，同样可采用中心极限定理来生成负二项分布随机数，其过程是：

（1）生成[0，1]均匀分布随机数 u；

（2）计算得到相应的标准正态分布随机数 z；

（3）计算$\dfrac{k(1-p)}{p}+z\sqrt{\dfrac{k(1-p)}{p^2}}$；

（4）按照四舍五入的原则，选取与$\dfrac{k(1-p)}{p}+z\sqrt{\dfrac{k(1-p)}{p^2}}$最接近的非负整数，并以此作为负二项分布随机数 x。

13.3.4 混合型随机变量的模拟

如果一个分布函数 $F_X(x)$ 既有连续部分又有离散部分，则其对应的随机变量为混合型随机变量。它的随机数 x 的生成方法为：如果[0，1]均匀分布的随机数 u 落在值域的严格连续递增的部分，则解方程 $F_X(x)=u$；如果 u 落在分布函数“跳”的区域，则随机数 x 取跳发生的点。

【例 13－10】 根据[0，1]区间上均匀分布 U 的随机数来表示如下密度函数：

$$f(x)=\begin{cases}p, & x=0\\(1-p)\lambda e^{-\lambda x}, & x>0\end{cases}$$

的随机变量的随机数。

解： 当 $u<p$ 时，$x=0$；

当 $u>p$ 时，$u=p+(1-p)(1-e^{-\lambda x})$，解得 $x=-\frac{1}{\lambda}\ln\left(\frac{1-u}{1-p}\right)$。

【例 13－11】 X 的密度函数 $f(x)$ 含有以下性质：(1) $f(x)=0$，$x<0$；(2) $f(0)=0.75$；(3) $f(x)=x^3$，$0<x<1$。根据[0，1]区间上均匀分布 U 的随机数列 0.1，0.7，0.8 产生 X 的随机数。

解： 由题知，$F(0)=0.75$，$F(x)=0.25x^4+c$，$0<x<1$，易得出 $c=0.75$。

根据 $X=F_X^{-1}(U)$，容易得出 $x_1=0$，$x_2=0$，$x_3=\sqrt[4]{4(0.8-0.75)}=0.6687$。

13.3.5 复合型随机变量的模拟

保险风险模型中的总损失常常需要采用复合分布的形式进行描述，对于由承保特征比较接近的保单构成的某类保单组合而言，若索赔次数用离散型随机变量 N 表示、每次索赔的金额用随机变量 X_i 表示，则索赔总量就可以看做是一个复合型随机变量。最常用的复合型随机变量是复合泊松分布和复合负二项分布的随机变量。在第 6 章中已对它们作过专门讨论，这里仅从随机模拟的角度讨论它们的分布。

若一个保险期（如一年）内的索赔总损失可以表示为：

$$S=\begin{cases}0, & N=0\\\sum_{i=1}^{N}X_i, & N>0\end{cases}$$

而且 X_1，$X_2\cdots$独立同分布，若 N 服从泊松分布，则称 S 为复合泊松分布；若 N 服从负二项分布，则称 S 服从复合负二项分布。

实际上，有了前面关于连续型随机变量和离散型随机变量的模拟方法，可以很自然地得到复合变量的模拟结果。设索赔次数 N 服从泊松分布，每

次索赔 X_i 服从相同的分布，用函数 $F(x)$ 表示，要产生复合泊松分布 S 的随机样本，一个简单方法就是对 N 和对 X_i 的分布分别进行模拟，比如首先生成泊松分布的随机数 n_1，再生成 n_1 个服从分布 $F(x)$ 的随机数 x_{11}，x_{12}，…，x_{1n_1}，令 $s_1 = \sum_{i=1}^{n_1} x_{1i}$，则 s_1 为复合泊松分布 S 的一个样本值。

复合负二项分布的模拟也可类似地进行。

13.3.6 正态随机向量的模拟

前面讨论的所有模拟都是对随机变量背景进行的，但是，现实中常常会出现随机向量的情景，这部分将对这个问题给出一些方法。下面将说明假设已经模拟得到标准 n 元正态随机数，则可以通过线性组合得到任意的 n 元正态随机数。

记 n 维($n \geqslant 2$)随机向量 $X = (X_1, \cdots, X_n)^T$，其均值为 $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^T$，协方差矩阵记做 Ω，Ω 的第 i 行 j 列元素为 $\sigma_{ij}^2 = E[(X_i - \mu_i)(X_j - \mu_j)]$，$i, j = 1, 2, \cdots, n$，即 X 服从分布 $N(\mu, \Omega)$。由 Ω 的定义知协方差矩阵 Ω 为正定矩阵，根据线性代数中的 Choleski 分解知，正定对称矩阵 Ω 可以表示为两个相同的非奇异矩阵 L 的乘积，即 $\Omega = L^T L$。其中，

$$L^T = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{12} & l_{22} & \cdots & 0 \\ \vdots & & & \vdots \\ l_{1n} & l_{2n} & \cdots & l_{nn} \end{bmatrix}$$

当 Ω 的分解式中的 L 为对角矩阵时，可证明 $X_1, \cdots, X_n$ 是相互独立的随机变量。

设 $Z = (Z_1, \cdots, Z_n)^T$ 是 n 元标准正态随机向量，服从分布 $N(0, I)$。因为正态随机变量的线性组合仍是正态随机变量，而且

$$E(LZ + \mu) = LE(Z) + \mu,\ Var(LZ + \mu) = LVar(Z)L^T = LL^T$$

可知 $W = LZ + \mu$ 服从分布 $N(\mu, \Omega)$。所以产生 $N(\mu, \Omega)$ 随机向量的算法如下：

(1) 产生 n 个独立同分布的标准正态随机变量，以向量 $z = (z_1, z_2, \cdots, z_n)^T$ 形式记录；

(2) 按公式 $x = \mu + L \cdot z$ 计算 n 维随机向量 $x = (x_1, x_2, \cdots, x_n)^T$，即：

$$x_j = \mu_j + \sum_{k=1}^{j} l_{jk} z_k, \qquad j = 1, 2, \cdots, n$$

【例 13－12】 某公司的资产分布于股票、债券和固定资产几个部分。高级管理层希望了解公司次年资产总额的情况，显然上述三类资产次年的价值是随机的，所以以随机向量 $X = (X_1, X_2, X_3)^T$ 的三个分量分别代表公

司当前的资产组合中股票、债券和固定资产的次年价值。研究表明，这三个随机变量的均值和协方差矩阵（以百万元为单位）分别为：

$$\mu=(3,1,1.5)^T,\quad \Omega=\begin{pmatrix}4 & -0.1 & 0.4\\ -0.1 & 2 & -0.21\\ 0.4 & -0.21 & 1\end{pmatrix}$$

试采用多元正态分布的随机模拟模型计算公司次年的资产总值超过 700 万元的概率。

解： 这里用 100 次模拟试验来解决该问题。为了说明模拟过程，下面用某次假想的试验进行说明，这里选用的数据是为简化计算方便说明，而非实际的随机数。首先计算三角矩阵 L：

$$L^T=\begin{pmatrix}2 & 0 & 0\\ -0.05 & \sqrt{2} & 0\\ 0.2 & -0.1\sqrt{2} & 0.2\end{pmatrix}$$

接下来考虑三个标准正态分布随机数构成的随机向量：

$$z=(z_1,z_2,z_3)^T=(1,-0.5,-1)^T$$

这里的数字进行了适当的简化，以便于简化计算。则有

$$x=\mu+L\cdot z=\begin{pmatrix}3\\1\\1.5\end{pmatrix}+\begin{pmatrix}2 & -0.05 & 0.2\\ 0 & \sqrt{2} & -0.1\sqrt{2}\\ 0 & 0 & 0.2\end{pmatrix}\begin{pmatrix}1\\-0.5\\-1\end{pmatrix}=\begin{pmatrix}4.825\\1-0.4\sqrt{2}\\1.3\end{pmatrix}$$

故这次试验得到的次年资产总额为：

$$\begin{pmatrix}4.825\\1-0.4\sqrt{2}\\1.3\end{pmatrix}^T\begin{pmatrix}1\\1\\1\end{pmatrix}\approx 6.559$$

近似为 660 万元。

100 次模拟结果列在表 13－6 中，并对结果作了升序排列。通过对这些试验结果进行分析，我们发现从编号为 67（结果用黑色标出）的试验开始资产总额超过了 700 万元，这表明在 100 次试验中有 34 次试验的总额超过了 700 万元，因此所估计的概率为 0.34。

表 13－6　　例 13－12 的计算结果　　单位：百万元

试验编号	股票	债券	固定资产	总额	试验编号	股票	债券	固定资产	总额
1	1.00	0.82	0.47	2.29	51	3.11	2.57	0.15	5.83
2	0.43	1.71	0.28	2.42	52	1.69	2.12	2.08	5.89
3	1.71	0.42	0.41	2.54	53	3.80	0.92	1.22	5.94
4	0.09	1.49	1.24	2.82	54	2.75	1.35	1.90	6.00
5	1.02	1.17	0.68	2.87	55	2.00	0.91	3.17	6.08

续表

试验编号	股票	债券	固定资产	总额	试验编号	股票	债券	固定资产	总额
6	1.34	1.46	0.23	3.03	56	2.66	2.12	1.36	6.14
7	1.59	0.24	1.25	3.08	57	3.30	0.09	2.81	6.20
8	0.56	1.41	1.23	3.20	58	3.02	2.82	0.46	6.30
9	1.60	0.44	1.25	3.29	59	3.59	0.40	2.51	6.50
10	2.61	0.34	0.42	3.37	60	4.48	0.77	1.37	6.62
11	0.08	1.52	1.98	3.58	61	4.93	0.73	0.97	6.63
12	2.25	0.44	1.02	3.71	62	3.58	2.83	0.38	6.79
13	1.90	0.68	1.17	3.75	63	2.70	0.86	3.26	6.82
14	1.38	0.85	1.66	3.89	64	4.54	1.62	0.69	6.85
15	1.81	0.86	1.24	3.91	65	3.34	2.24	1.32	6.90
16	1.02	1.90	1.07	3.99	66	2.87	2.98	1.10	6.95
17	1.10	1.74	1.28	4.12	67	4.21	1.94	0.93	7.08
18	1.46	2.21	0.47	4.14	68	0.71	3.68	2.83	7.22
19	3.06	0.61	0.69	4.36	69	2.86	1.99	2.38	7.23
20	2.94	1.18	0.25	4.37	70	4.79	0.62	1.87	7.28
21	2.47	0.61	1.37	4.45	71	3.76	2.62	1.09	7.47
22	1.37	0.56	2.59	4.52	72	3.12	1.98	2.42	7.52
23	0.71	3.70	0.16	4.57	73	4.53	0.60	2.43	7.56
24	0.81	2.53	1.31	4.65	74	4.38	3.22	0.01	7.61
25	0.85	0.18	3.65	4.68	75	3.45	3.71	0.48	7.64
26	0.00	3.66	1.03	4.69	76	2.92	2.73	2.22	7.87
27	1.21	0.41	3.10	4.72	77	3.67	1.24	2.99	7.90
28	2.15	1.18	1.39	4.72	78	4.59	2.84	0.50	7.93
29	2.94	0.59	1.22	4.75	79	5.32	2.44	0.40	8.16
30	1.67	2.02	1.10	4.79	80	3.24	3.27	1.65	8.16
31	4.20	0.62	0.00	4.82	81	4.34	1.89	2.04	8.27
32	1.54	1.72	1.57	4.83	82	6.12	0.28	2.18	8.58
33	0.42	2.87	1.59	4.88	83	6.00	0.44	2.29	8.73
34	3.03	0.29	1.57	4.89	84	5.30	0.23	3.27	8.80
35	3.76	0.47	0.83	5.06	85	5.43	0.43	3.02	8.88
36	2.46	0.76	1.87	5.09	86	4.77	1.04	3.18	8.99
37	2.78	0.01	2.36	5.15	87	6.08	0.40	2.58	9.06
38	2.87	0.83	1.46	5.16	88	4.43	3.23	1.62	9.28
39	2.65	2.50	0.02	5.17	89	4.76	2.62	1.96	9.34
40	3.13	1.43	0.62	5.18	90	4.11	3.62	1.68	9.41

续表

试验编号	股票	债券	固定资产	总额	试验编号	股票	债券	固定资产	总额
41	1.24	0.82	3.18	5.24	91	6.12	0.56	3.49	10.17
42	0.24	3.94	1.14	5.32	92	5.86	1.60	2.74	10.20
43	3.22	0.90	1.20	5.32	93	7.75	1.47	1.06	10.28
44	3.49	1.37	0.47	5.33	94	4.99	2.31	3.16	10.46
45	2.17	1.12	2.06	5.35	95	7.03	0.32	3.44	10.79
46	2.53	0.00	2.88	5.41	96	5.60	1.15	4.05	10.80
47	2.34	1.62	1.51	5.47	97	5.13	3.45	2.45	11.03
48	2.73	1.11	1.89	5.73	98	6.59	4.13	0.36	11.08
49	3.73	1.02	0.99	5.74	99	7.50	1.65	2.08	11.23
50	0.40	4.29	1.14	5.83	100	9.11	0.13	4.19	13.43

§13.4　模拟样本的容量

模拟实验次数或模拟样本容量是影响随机模拟成败的一个重要因素。模拟样本的大小取决于概率分布的形式和对估计值的精确程度的要求。一般而言，对估计值的精确度要求越高，对样本容量要求就越大。下面我们举例说明。

假定我们要估计贝努里随机变量的成功概率（或均值）p，得到了容量为 n 的随机样本，并且以样本的成功比率 $\hat{p}_n$ 作为 p 的估计值。这时可以考虑以下两种基本问题：首先，给定 n 和 α，$0<\alpha<1$，求满足 $P\{|\hat{p}_n-p|<\varepsilon\}=1-\alpha$ 的 ε 值；其次，给定 ε 和 α，求满足上式所需的样本容量 n。换而言之，在第一类问题中，对给定的样本容量确定估计值的精确度；在第二类问题中，对给定的估计精度，确定必要的样本容量。

特别地，取 $\alpha=0.01$，$n=100$，并假定 p 的真值为 0.5（$p=0.5$ 时得到的方差最大），求相应的 ε 值。这里可以将 n 看做“充分大”，由中心极限定理可得：

$$\frac{\hat{p}_n-p}{\sqrt{\dfrac{(1-p)p}{n}}}=\frac{\hat{p}_n-p}{0.05}$$

可以用标准正态分布来近似。由正态分布函数表可知：

$$P\left[-2.58<\frac{\hat{p}_n-p}{0.05}<2.58\right]=0.99$$

即　$P[-0.129<\hat{p}_n-p<0.129]=P[|\hat{p}_n-p|<0.129]=0.99$

故 $\varepsilon=0.129$。

若取 $\alpha=0.02$，$\varepsilon=0.01$，通过类似的方法可得到需要的样本容量 n 约为 13 526。

在一般的随机模拟问题中都会事先对估计值的精确程度作出规定，然后按照精度确定相应的模拟样本容量，这样得到的样本容量或试验次数值一般都很大（如上面的 13 526），因而用手工计算进行模拟几乎是不可能的，必须进行计算机操作。

【例 13－13】 假设随机变量 X 的均值为 μ，方差为 σ^2。通过模拟来估计 μ。

（1）求使得真值与估计值的相对误差不大于 5% 的概率为 0.9 所需的模拟次数；

（2）求使得真值与估计值的相对误差不大于 1% 的概率为 0.9 所需的模拟次数；

（3）求使得真值与估计值的相对误差不大于 1% 的概率为 0.95 所需的模拟次数。

解：（1）由题得，对估计值的要求为 $P\left(\frac{|\bar{X}-\mu|}{\mu}\leqslant 0.05\right)=0.9$

可得：$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\leqslant\frac{0.05\mu\sqrt{n}}{\sigma}\right)=0.95$，　　$\frac{0.05\mu\sqrt{n}}{\sigma}=\Phi^{-1}(0.95)=1.645$

所以，$n=\left(\frac{1.645}{0.05}\right)^2\left(\frac{\sigma}{\mu}\right)^2$。

（2）类似地，可求使得真值与估计值的相对误差不大于 1% 的概率为 0.9 所需的模拟次数为：

$$n=\left(\frac{1.645}{0.01}\right)^2\left(\frac{\sigma}{\mu}\right)^2$$

（3）使得真值与估计值的相对误差不大于 1% 的概率为 0.95 所需的模拟次数为：

$$n=\left(\frac{1.96}{0.01}\right)^2\left(\frac{\sigma}{\mu}\right)^2$$

【例 13－14】 在一个繁忙的交通路口，有大量汽车经过。设每辆车在一天的某段时间内出事故的概率为 0.001。某交通部门希望每天该路口出事故次数小于等于 2 的概率为 0.95，那么每天在这个时间段内该路段限车辆数为多少？

解：令 $X_i=\begin{cases}1，\text{第 } i \text{ 辆汽车发生事故}\\0，\text{第 } i \text{ 辆汽车没发生事故}\end{cases}$，则

$$E(X_i)=0.001，\quad Var(X_i)=0.001\times 0.999$$

令 $\bar{X}=\frac{\sum_{i=1}^{n}X_i}{n}$，则 $\frac{\bar{X}-0.001}{\sqrt{0.001\times 0.999/n}}$ 近似服从标准正态分布。

若要满足 $P(\sum_{i=1}^{n} X_i \leqslant 2) = 0.95$，则

$$P(\sum_{i=1}^{n} X_i \leqslant 2) = P(\sum_{i=1}^{n} X_i/n \leqslant 2/n)$$

$$= P\left(\frac{\sum_{i=1}^{n} X_i/n - 0.001}{\sqrt{0.001 \times 0.999/n}} \leqslant \frac{\frac{2}{n} - 0.001}{\sqrt{0.001 \times 0.999/n}}\right) = 0.95$$

所以，$\dfrac{\frac{2}{n} - 0.001}{\sqrt{0.001 \times 0.999/n}} = 1.645$，可得出 $n = 662$。所以交通部门在此时间段应限量通行汽车 662 辆。

在随机模拟中，常采用方差缩减技术（Variance Reduction Techniques）来提高随机模拟的效率。该方法利用已知信息来减少在给定精度下需要的样本容量，或对给定的样本容量提高估计精度。

§13.5 Bootstrap 模拟

13.5.1 引　言

我们考虑一个问题：假设随机变量 X_1，$X_2, \cdots, X_n$ 是独立同分布于 $F(x)$，且已知一组样本观测值 $x_1, x_2, \cdots, x_n$ 和函数 $g(x_1, x_2, \cdots, x_n)$。如何计算 $Var(g(X_1, X_2, \cdots, X_n))$？

如果分布函数 $F(x)$ 已知，这个问题容易解决。一方面可以利用分布函数直接计算 $Var(g(X_1, X_2, \cdots, X_n))$ 的值。另一方面，只要能产生服从分布 $F(x)$ 的充分多的随机数，再将这些模拟数代入函数 g 中，可得与 $Y = g(X_1, X_2, \cdots, X_n)$ 同分布的充分多的随机数，则理论上就能得到 Y 的分布 $F_Y(y)$ 的任何信息，从而得到 $Var(Y)$ 的估计值。如果函数 $F(x)$ 的具体形式未知，则 Y 的分布未知。若有多组模拟样本 $(x_{11}, x_{12}, \cdots, x_{1n})$，$\cdots$，$(x_{m1}, x_{m2}, \cdots x_{mn})$ 共 mn 个，那么对于每个模拟样本都可以产生具体的 y，生成 y_1，$\cdots y_m$，计算其样本方差，则可以得到 $Var(g(X_1, X_2, \cdots, X_n))$ 的近似估计值。

实际中，很难获得多组观测样本，往往只有一组样本观测值 x_1，x_2，$\cdots, x_n$，因此能计算一个 $g(x_1, x_2, \cdots, x_n)$ 的值，无法计算 $Var(g(X_1, X_2, \cdots, X_n))$。在只有一组样本和对总体信息一无所知的情况下，自助法能产生多组模拟样本，给出各种样本统计量的均值和方差一个较好的估计。

13.5.2 Bootstrap 模拟的思想

尽管分布函数 $F(x)$ 未知，但是当我们观测到 n 个数据点后，很容易判

断此分布看起来像什么分布。由 Glivenko-cantelli 定理知，当 $n\to\infty$，经验分布 $F_n(x)$ 以概率 1 收敛到真实分布 $F(x)$。故当 n 充分大时，$F_n(x)$ 应当接近于 $F(x)$。因此可以认为，$F_n(x)$ 是随机变量 X 的近似分布函数。从而在一定意义下，

$$Var_{F_n}(g(X_1,\cdots,X_n))=E_{F_n}(g(X_1,\cdots,X_n)-E_{F_n}(g(X_1,\cdots,X_n))^2$$

可以作为 $Var(g(X_1,X_2,\cdots,X_n))$ 的近似，这里 E_{F_n}，Var_{F_n} 表示期望依赖于经验分布 $F_n(x)$。这个式子中的 X_i 应当理解为相互独立，且分布同 F_n。

要计算 $Var_{F_n}(g(X_1,\cdots,X_n))$，需要对经验分布抽取所有可能的样本，求出 $g(x_{i_1},x_{i_2},\cdots,x_{i_n})$，然后得到样本估计量的样本方差。对于一个等概率 n 点离散分布，即 $P(X=x_i)=1/n$，$i=1,2,\cdots,n$，$Var_{F_n}(g(X_1,\cdots,X_n))$ 的求解有两种方法：

（1）当 n 比较小时，$Var_{F_n}(g(X_1,\cdots,X_n))$ 常常是可以得到精确值的。因为若 $X_1,X_2,\cdots,X_n$ 是独立同分布于等概率 n 点离散分布，则类似一个将 n 面的骰子投掷 n 次，所有可能的组合有 n^n 种，事实上它们的联合分布是一个多项分布，所以，

$$\begin{aligned}Var_{F_n}(g(X_1,\cdots,X_n)) &= E_{F_n}[(g(X_1,X_2,\cdots,X_n)-\bar{g})^2]\\ &=\sum_{i_n=1}^{n}\cdots\sum_{i_1=1}^{n}(g(x_{i_1},x_{i_2},\cdots,x_{i_n})-\bar{g})^2\\ &\quad P(X_1=x_{i_1},\cdots,X_n=x_{i_n})\\ &=\sum_{i_n=1}^{n}\cdots\sum_{i_1=1}^{n}\frac{(g(x_{i_1},x_{i_2},\cdots,x_{i_n})-\bar{g})^2}{n^n}\end{aligned}$$

这里 $\bar{g}=\dfrac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}g(x_{i_1},\cdots,x_{i_n})$。

（2）当 n 很大时，n^n 项求和不方便。根据大数定理，对任意随机变量 X，只要能生成独立与 X 同分布的随机变量序列 $\{\xi_i\}$，则当 m 很大时，$E(f(X))\approx\dfrac{1}{m}\sum_{i=1}^{m}f(\xi_i)$。这就是所谓的模拟方法。因为产生一个服从经验分布函数 F_n 的随机数相当于将 n 面的均匀骰子投掷一次，这是很容易进行的，所以可以做 m 次试验，每次试验将 n 面的均匀骰子投掷 n 次，假设第 i 次试验的结果为 $x_1^i,x_2^i,\cdots,x_n^i$，计算 $y_i=g(x_1^i,\cdots,x_n^i)$。把

$$\widehat{Var}(Y)=\frac{1}{m-1}\sum_{i=1}^{m}(g(x_1^i,x_2^i,\cdots,x_n^i)-\bar{y})^2$$

$$\bar{y}=\frac{1}{m}\sum_{i=1}^{m}g(x_1^i,x_2^i,\cdots,x_n^i)$$

为 $Var_{F_n}(g(X_1,\cdots,X_n))$ 的近似值，也作为 $Var(g(X_1,X_2,\cdots,X_n))$ 的近似值。

例如，要计算 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 的方差 $Var(s^2)$，其中 $\{X_i\}$ 是独立同分布于 $F(x)$ 的随机变量，则可做 k 组模拟，每组模拟产生 n 个 $F_n(x)$ 的随机数，每组模拟数可以计算出一个 s^2 的观测值，记为 $s^2(i)$，$i=1,\cdots,k$，可得这 k 个 s^2 的观测值的样本方差为 $\widehat{Var}(s^2)=\frac{1}{k-1}\sum_{j=1}^{k}[s^2(i)-\frac{1}{k}\sum_{j=1}^{k}s^2(i)]^2$。当 k 和 n 充分大时，可证明 $\widehat{Var}(s^2)$ 可以非常接近统计量 s^2 的理论方差 $Var(s^2)$。

这种以 x_1，$x_2,\cdots,x_n$ 为新总体，$F_n(x)$ 为总体分布，进行随机抽样，产生样本的方法称为 Bootstrap 法。将 n 面的骰子投掷 n 次称为一次试验，产生一个自助样本，m 个自助样本相当于投掷 mn 次。每次试验的投掷次数总是和原始观测值个数相等。由于自助法是对原始观测值做有放回的随机试验，所以同一个观测值可能多次出现，可能会一次也不出现。自助法依赖于原始观测值，原始观测值不同，相当于骰子不同。

值得注意的是，自助法实际上是对有限离散分布的模拟，它是在原始样本的基础上继续模拟样本得到的，并没有考虑到总体的所有的观测值。这样求出的方差只能是 $Var(g(X_1,X_2,\cdots,X_n))$ 的近似值。

【例 13　15】 假设随机变量 X_1，X_2，X_3 是独立且同分布于 $F(x)$，如果用样本方差 s^2 作为总体方差 σ^2 的估计，即 $\hat{\sigma}^2=\frac{\sum_{i=1}^{3}(x_i-\bar{x})^2}{3}$。假设有一个样本原始观测值为 1，2，3，则样本方差的观测值为$\frac{2}{3}$，估计 $Var(\hat{\sigma}^2)$。

解： 如果需要评价统计量 $\hat{\theta}$ 的估计好坏，在没有其他信息时，这是无法进行的，但自助法可以给出一个数值估计。

如果对原始数据进行 3 次模拟，假设结果为{1，2，3}，{1，2，2}，{1，3，3}，其对应的方差为：2/3，2/9，8/9，所以这 3 个数可以得到一个均值为 16/27，相应的波动为：

$$\frac{1}{3}\times\left[\left(\frac{2}{3}-\frac{16}{27}\right)^2+\left(\frac{2}{9}-\frac{16}{27}\right)^2+\left(\frac{8}{9}-\frac{16}{27}\right)^2\right]=\frac{56}{729}$$

这个结果可以作为 σ^2 的估计量 $\hat{\sigma}^2$ 的波动性能指标值，即 $Var(\hat{\sigma}^2)$ 的近似估计值。■

总的来说，自助法的基本思想是：（1）经验分布渐近原分布，所以来自经验分布的统计量可以近似来自原分布的统计量。（2）只要能从分布中生成足够多的模拟数，则随机模拟可以充分描述分布的任何信息。（3）经验分布是有限总体的离散分布，易于进行随机模拟。所以自助法是原始观测值做有放回的随机试验，用试验产生的多组数据估计想要的方差或其他

的量。其中包含了两步近似，一是随机模拟近似经验分布的概率性质，二是经验分布渐近原分布，所以自助法的精度既和模拟次数有关，也和原始观测值个数有关。

13.5.3 Bootstrap 计算均方误差

假设随机变量 X_1，$X_2,\cdots,X_n$ 是独立同分布 $F(x)$，$\theta(F)$是分布 $F(x)$的参数，假设 $\theta(F)$的一个估计量为 $\hat{\theta}=g(X_1, X_2, \cdots, X_n)$，为了判别估计量的好坏，采用均方误差作为判别标准，即

$$MSE(F_n)=E_{Fn}[(g(X_1, X_2, \cdots, X_n)-\theta(F_n))^2]$$
$$=Var_F(g(X_1, X_2, \cdots, X_n))+[Bias(g(X_1, X_2, \cdots, X_n))]^2$$

其中 $Bias(g(X_1, X_2, \cdots, X_n))=\theta-E_F g(X_1, X_2, \cdots, X_n)$，$E_F$ 表示期望依赖于分布 $F(x)$。

如果 $g(X_1, X_2, \cdots, X_n)$是 $\theta(F)$的无偏估计，则 $MSE(F)=Var_F(g(X_1, X_2, \cdots, X_n))$。如果 $g(X_1, X_2, \cdots, X_n)$不是 $\theta(F)$的无偏估计，则

$$MSE(F)=E_F[(g(X_1, X_2, \cdots, X_n)-\theta(F))^2]$$

当 $F(x)$未知或是很复杂时，理论上 $MSE(F)$无法得到。这时可用

$$MSE(F_n)=E_{F_n}[(g(X_1, X_2, \cdots, X_n)-\theta(F_n))^2]$$
$$=\sum_{i_n=1}^{n}\cdots\sum_{i_1=1}^{n}(g(x_{i_1}, x_{i_2}, \cdots, x_{i_n})-\theta(F_n))^2$$
$$P(X_1=x_{i_1}, \cdots, X_n=x_{i_n})$$
$$=\sum_{i_n=1}^{n}\cdots\sum_{i_1=1}^{n}\frac{(g(x_{i_1}, x_{i_2}, \cdots, x_{i_n})-\theta(F_n))^2}{n^n}$$

或者自助法产生模拟样本 $x_1^i,\cdots,x_n^i$，$1\leqslant i\leqslant m$，计算

$$MSE(F_n)\approx\frac{1}{m}\sum_{i=1}^{m}(g(x_1^i, x_2^i, \cdots, x_n^i)-\theta(F_n))^2$$

作为 $MSE(F)$的近似。后者称为均方误差 $MSE(F)$的自助法近似值。

【例 13－16】 假设随机变量 X_1，X_2，X_3 是独立同分布 $F(x)$，如果用样本方差作为总体方差 σ^2 的估计，即 $\hat{\theta}=\dfrac{\sum_{i=1}^{3}(x_i-\bar{x})^2}{3}$。假设有一个样本原始观测值为 1，2，3，要求：

（1）求 $MSE(F_n)$；

（2）如果对原始数据进行 3 次模拟，假设结果为{1，2，3}，{1，2，2}，{1，3，3}，运用自助方法，计算均方误差的自助法近似值。

解：

（1）由于只有 3 个观测值，需要做 3^3 次试验，每次试验将 3 面的均匀骰子投掷 3 次，得到结果如下：即以 1/27 的概率得到{(1，1，1)、(2，2，

2)、(3，3，3)}，模拟样本方差依次为{0，0，0}；以 1/9 的概率得到{(1，1，2)、(1，1，3)、(2，2，3)、(2，2，1)、(3，3，1)、(3，3，2)}，模拟样本方差依次为$\left\{\frac{2}{9},\frac{8}{9},\frac{2}{9},\frac{2}{9},\frac{8}{9},\frac{2}{9}\right\}$，以$\frac{2}{9}$的概率得到{(1，2，3)}，方差为$\frac{2}{3}$。

$$\theta(F_n)=\frac{1}{3}[(1-2)^2+(2-2)^2+(3-2)^2]=\frac{2}{3}$$

$$\begin{aligned}MSE(F_n)&=\sum_{i_n=1}^{n}\cdots\sum_{i_1=1}^{n}\frac{(g(x_{i_1},x_{i_2},\cdots,x_{i_n})-\theta(F_n))^2}{n^n}\\&=\frac{1}{27}\left[\left(0-\frac{2}{3}\right)^2+\left(0-\frac{2}{3}\right)^2+\left(0-\frac{2}{3}\right)^2\right]\\&\quad+\frac{1}{9}\left[\left(\frac{2}{9}-\frac{2}{3}\right)^2\times4+\left(\frac{8}{9}-\frac{2}{3}\right)^2\times2\right]+0\\&=\frac{4}{27}\end{aligned}$$

(2) 这 3 个自助样本{1，2，3}，{1，2，2}，{1，3，3}的方差依次为 2/3，2/9，8/9。

$$\begin{aligned}MSE(F_n)&\approx\frac{1}{m}\sum_{i=1}^{m}(g(x_1^i,x_2^i,\cdots,x_n^i)-\theta(F_n))^2\\&=\frac{1}{3}\left[\left(\frac{2}{3}-\frac{2}{3}\right)^2+\left(\frac{2}{9}-\frac{2}{3}\right)^2+\left(\frac{8}{9}-\frac{2}{3}\right)^2\right]\\&=\frac{20}{243}\end{aligned}$$

■

读者可比较例 13－16 与例 13－15 的区别。

【例 13－17】 假设随机变量 X_1，$X_2,\cdots,X_{15}$是独立同分布 $F(x)$，现有它们的一组观测值：8.0、5.1、2.2、8.6、4.5、5.6、8.1、6.4、3.3、7.3、8.0、4.0、6.5、6.3、9.1。下面是被模拟的经验分布的自助样本，每个样本有 15 个数据：

样本 1： 3.3、3.3、3.3、4.0、4.0、4.0、6.5、7.3、8.1、8.6、8.8、9.1、9.1、9.1、9.1

样本 2： 3.3、3.3、5.1、5.1、5.1、5.6、6.4、6.5、6.5、7.3、7.3、8.0、8.0、8.0、8.1

样本 3： 2.2、3.3、3.3、4.0、4.5、4.5、5.1、6.5、7.3、8.0、8.1、8.1、8.1、9.1、9.1

(1) 假设用样本均值 $\bar{X}$ 估计分布的均值 θ，使用这三个自助样本来估计 $\bar{X}$ 的 MSE；

(2) 若 a 是样本方差，b 是经验分布的方差，求 $a-b$；

(3) 假设 $\theta = F(5)$，如果用 $\hat{\theta} = \frac{\#\{i: X_i \leq 5\}}{15}$，使用这 3 个自助样本来估计 $\hat{\theta}$ 的 MSE。

解：

(1) 因为 $\theta(F_n) = E_{F_n}(X)$

$= (8.0 + 5.1 + 2.2 + 8.6 + 4.5 + 5.6 + 8.1 + 6.4 + 3.3 + 7.3 + 8.0 + 4.0 + 6.5 + 6.3 + 9.1)/15 = 6.2$

而 3 个自助样本的样本均值分别是：6.51，6.24，6.08。所以 $MSE(F)$ 的自助法近似值为：

$$\frac{1}{3}[(6.51 - 6.20)^2 + (6.24 - 6.20)^2 + (6.08 - 6.20)^2] = 0.037$$

(2) 样本方差是：

$$S_{15}^2 = \frac{1}{14}\sum_{i=1}^{15}(x_i - \bar{x})^2 = \frac{1}{14}[\sum_{i=1}^{15}x_i^2 - 15(\bar{x})^2]$$

$$= \frac{1}{14}[635.92 - 15(6.2)^2] = 4.237$$

经验分布的方差是：

$$Var_{F_n}[X] = \frac{1}{15}\sum_{i=1}^{15}(x_i - \bar{x})^2 = \frac{1}{15}[\sum_{i=1}^{15}x_i^2 - 15(\bar{x})^2] = \frac{14}{15}S_{15}^2 = 3.955$$

$$a - b = S_{15}^2 - \frac{14}{15}S_{15}^2 = \frac{1}{15}S_{15}^2 = 0.28$$

(3) 因为 $\theta(F) = F(5)$，用经验分布替换 $F(x)$，得

$$\theta(F_n) = F_n(5) = \frac{\#\{i: x_i \leq 5\}}{15}$$

$$= \frac{\#\{(8.0, 5.1, 2.2, 8.6, 4.5, 5.6, 8.1, 6.4, 3.3, 7.3, 8.0, 4.0, 6.5, 6.3, 9.1) \leq 5\}}{15}$$

$$= 0.267$$

又因为对于自助样本有

$$\hat{\theta}(x_1^1, x_2^1, \cdots, x_n^1)$$

$$= \#\{(3.3, 3.3, 3.3, 4.0, 4.0, 4.0, 6.5, 7.3, 8.1, 8.6, 8.8, 9.1, 9.1, 9.1, 9.1) < 5\}/15 = \frac{6}{15} = 0.4$$

$$\hat{\theta}(x_1^2, x_2^2, \cdots, x_n^2) = \frac{2}{15} = 0.133$$

$$\hat{\theta}(x_1^3, x_2^3, \cdots, x_n^3) = \frac{6}{15} = 0.4$$

根据式 (13.5.1) 得：

$$MSE(F_e) \approx \frac{1}{3}\sum_{i=1}^{3}(g(x_1^i, x_2^i, \cdots, x_n^i) - \theta(F_e))^2$$

$$= \frac{1}{3}[(0.4-0.267)^2 + (0.133-0.267)^2 + (0.4-0.267)^2]$$

$$= 0.0178$$

■

§13.6 MCMC 模拟

假设某个函数 $h(x)$ 关于概率分布 $\pi(x)$ 的期望 $E_\pi(h(X)) = \int h(x)\pi(x)dx$ 存在。如果能够生成服从 $\pi(x)$ 的 n 个独立的样本 $x_1, \cdots, x_n$，根据大数定律，对充分大的 n，可以取 $\hat{y}_n = \sum_{k=1}^{n} h(x_k)/n$ 作为积分 $E_\pi(h(X))$ 的近似值，这便是蒙特卡罗模拟的基本思想。前面已经介绍了线性同余法、反函数法、取舍法等多种产生服从特定分布的随机数的方法。但是这些方法只适用于简单的分布函数，如指数分布、韦伯分布等。在大多数实际应用中，人们感兴趣的分布 $\pi(x)$ 一般高维、复杂，产生独立样本是不可行的。通常情况下，产生的样本或是相关的，或者产生的样本异于所要求的分布 $\pi(x)$。因此如何产生服从一个复杂、高维的分布 $\pi(x)$ 的随机样本便成了非常重要的问题。

米特罗波利斯（Metropolis）等人在 1953 年最早给出了通过生成一个马尔可夫链来实现从分布 $\pi(x)$ 中采样（生成相关的样本）这一重要思想。随后，哈斯汀（Hastings）将其推广到更一般的形式，其基本原理便是建立一个马尔可夫链，以 $\pi(x)$ 为平稳分布，从马尔可夫链中抽取样本。这类方法称为 **MCMC 方法**。

本节首先给出马尔可夫链的基本知识和性质，然后说明 MCMC 的原理，最后介绍两种常用的 MCMC 方法。

13.6.1 马尔可夫链

1. 定义。假设随机过程 X_t 表示为系统在时间 t 的状态，若在 X_t 已知的条件下，系统在将来时刻 X_{t+n} 的状态（或某些取值）的概率与过去状态 X_s $(s<t)$ 的状态无关，只与过程在 t 时刻的状态有关，则称随机过程 $\{X_t\}$ 为**马尔可夫过程**。设 S 表示 $\{X_t, t\geqslant 0\}$ 所处的状态空间，若 S 取离散值 $S=\{a_1, a_2, \cdots\}$，则称 $\{X_t\}$ 为**马尔可夫链**，简称**马氏链**。

当 $0\leqslant t_1 < t_2 \cdots < t_r < m$，$n>0$，$r>0$，由马氏链的定义知，

$$P_{ij}(m, m+n) = P\{X_{m+n}=a_j \mid X_{t_1}=a_{i_1}, X_{t_2}=a_{i_2}, \cdots X_{t_r}=a_{i_r}, X_m=a_i\}$$
$$= P\{X_{m+n}=a_j \mid X_m=a_i\} \qquad (13.6.1)$$

其中 $a_{i_k}\in S$，$1\leqslant k\leqslant n$，则称 $P_{ij}(m,m+n)=P\{X_{m+n}=a_j|X_m=a_i\}$ 为随机过程 X_t 在时刻 m 处于状态 a_i 的条件下，在时刻 $m+n$ 转移到状态 a_j 的转移概率。显然有

$$\sum_{j=1}^{\infty}P_{ij}(m,m+n)=1,\qquad i=1,2,\cdots$$

当一个马氏链的转移概率 $P_{ij}(m,m+n)$ 只与 i，j 及时间间隔 n 有关时，即

$$P_{ij}(m,m+n)=P(X_{m+n}=a_j|X_m=a_i)=P(X_n=a_j|X_0=a_i)\triangleq P_{ij}(n),$$

称转移概率具有时间齐次性，这时马氏链称为齐次马氏链。$P(n)=(P_{ij}(n))$ 为 n 步转移概率矩阵。一步转移矩阵 $P(1)$ 简记为 P，用如下的矩阵表示：

$$P(1)=(P_{ij})=\begin{pmatrix}P_{11}&P_{12}&\cdots&P_{1j}&\cdots\\P_{21}&P_{22}&\cdots&P_{2j}&\cdots\\\vdots&\vdots&\cdots&\vdots&\cdots\\P_{i1}&P_{i2}&\cdots&P_{ij}&\cdots\\\vdots&\vdots&\cdots&\vdots&\cdots\end{pmatrix}$$

容易证明，齐次马氏链的转移矩阵满足下面的 Chapman-Kolmogorov 方程：

$$\begin{aligned}P_{ij}(m+n)&=\sum_{k\in S}P_{ik}(m)P_{kj}(n)\\&=\sum_{k\in S}P_{ik}(n)P_{kj}(m)\end{aligned}$$

因此，$P(k)=\overbrace{P\cdots P}^{k}=P^k$。

2. *遍历性马氏链及平稳概率*。假设 $\{\pi_0(i)\}$ 为马氏链的初始分布，即 $\pi_0(i)=P\{X_0=a_i\}$，$a_i\in S$。记 $\pi_t=(\pi_t(1),\cdots\pi_t(n),\cdots)'$ 为 X_t 的分布，则

$$\pi_{t+1}(i)=\sum_{i\in S}P(X_{t+1}=a_j\mid X_t=a_i)P(X_t=a_i)=\sum_{j\in S}\pi_t(i)P_{ij}\tag{13.6.2}$$

从而

$$\pi'_{t+1}=\pi'_tP=\pi'_{t-1}P\cdot P=\cdots=\pi'_0P^t$$

若马氏链经过相当长的一段运行时间过程处于状态 j 的时间的比例（即 $\pi_\infty(j)=\lim\limits_{n\to\infty}P(X_n=j)$）只与状态 j 有关，而不论初始分布如何，则马氏链存在极限分布，记 $\{\pi_\infty(i)\}$ 为极限分布。若极限分布存在，则对式（13.6.2）两边取极限有：

$$\begin{aligned}&\pi_\infty(j)=\sum_i\pi_\infty(i)P_{ij},\qquad j=1,2,\cdots\\&\sum_i\pi_\infty(i)=1\end{aligned}\tag{13.6.3}$$

若马氏链的分布族 $\{\pi_t,t\geqslant0\}$ 满足 $\pi_{t+1}(i)=\pi_t(i)$，$\forall a_i\in S$，$t\geqslant0$，则

称马氏链是平稳马氏链。注意到 $\pi_1' = \pi_0'P$，因此若 π_0'满足

$$\pi_0'P = \pi_0' \tag{13.6.4}$$

则马氏链是平稳马氏链，π_0 称为平稳分布。因此，如果马氏链的极限分布存在，则 π_∞ 就是马氏链的平稳分布。显然，不管初始状态 X_t 是什么分布，在经过一段时间后，X 的边际分布就都是平稳分布 π_∞。

如果对于每一对状态 a_i 和 a_j，马氏链从状态 a_i 经过有限步后转移到状态 a_j 的概率大于 0，即存在 $n\geqslant 1$，$P_{ij}^{(n)}>0$，则称这样的马氏链是不可约的。对于不可约的马氏链，如果对某个 $n\geqslant 0$ 和某个状态 a_j 有

$$P(X_n = j | X_0 = j) > 0 \text{ 和 } P(X_{n+1} = j | X_0 = j) > 0$$

则称这个不可约的马氏链是遍历的。

对于遍历的马氏链，可以证明极限分布存在且唯一，转移概率具有极限分布 π_∞，即$\lim\limits_{n\to\infty} P_{ij}(n) = \pi_\infty(j)$（不依赖 i）。

具有遍历性的马氏链的一个重要性质是对于状态空间中的任意函数 f，

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \sum_j \pi_\infty(j) f(j) \tag{13.6.5}$$

以概率 1 成立。

遍历马氏链的充要条件不容易验证，这里给出一个充分条件：

若齐次马氏链$\{X_n, n\geqslant 0\}$的状态空间为 $S = \{a_1, a_2, \cdots, a_N\}$，存在正整数 m，使对任意的 a_i，$a_j \in S$，都有 $P_{ij}(m) > 0$，$i, j = 1, 2, \cdots, N$，则马氏链具有遍历性。

13.6.2　MCMC 基本原理

现在想要生成具有概率分布 $P(X = j) = \pi(j)$，$j = 1, \cdots, N$ 的随机变量 X 的随机数。如果能够生成一个具有极限概率$\{\pi(j), j = 1, \cdots N\}$的遍历马氏链$\{X_t\}$，则根据遍历马氏链的性质，不管初始状态 X_0 是什么分布，在经过一段时间后，$\{X_t\}$的边际分布近似于极限分布 π。因此我们通过运行该马氏链 n 步（n 足够大），来获得 X_n 的值，并将这个值近似的作为想要生成的随机变量随机数。另外，如果是为了能够估计 $E_\pi(h(X)) = \sum\limits_{j=1}^{N} h(j)\pi(j)$，可选取马氏链中的值 $X_1, \cdots, X_n$，可以使用估计量 $\frac{1}{n}\sum\limits_{i=1}^{n} h(X_i)$ 来估计$E_\pi(h(X))$，即 $\frac{1}{n}\sum\limits_{i=1}^{n} h(x_i) \to E_\pi(h(X))$。然而，由于初始状态的选取对前期的马氏链的状态有很强的影响，一般在应用中适当的选取某个 k，然后舍去前 k 个状态，也就是说，使用估计量 $\frac{1}{n-k}\sum\limits_{i=k+1}^{n} h(X_i)$。至于 k 的精

确选取，Aarts 和 Korst 给出了一些结果[1]。一般情况下，可以用直观给出的值（通常来说这样做会有比较好的效果，因为无论取什么样的值都能够有很好的收敛）。

MCMC 的一个关键问题是如何构造一个极限分布为 $\{\pi(j), j=1, \cdots, N\}$ 的遍历马氏链。这里介绍两种比较流行的方法。

13.6.3 Metropolis-Hastings 抽样

对于任意给定的概率分布 $\{\pi(x), x=1, \cdots, N\}$，Metropolis - Hastings 方法描述了一个马氏链的转移准则，使得其稳定分布为 $\pi(x)$。

设 Q 是个状态空间为整数 $\{1, \cdots, N\}$ 的不可约的转移概率矩阵，其第 x 行第 y 列元素为 $q(x, y)$，函数 $\alpha(x, y)$ 满足 $0 \leqslant \alpha(x, y) \leqslant 1$，$x, y \in \{1, \cdots, N\}$。对于任一组合 (x, y)，定义：

$$p(x, y) = q(x, y)\alpha(x, y) \qquad x \neq y$$

$$p(x, x) = 1 - \int_{x \neq y} q(x, y)\alpha(x, y)dy \qquad x = y$$

则易见 $p(x, y)$ 构成一个概率转移矩阵。

此方法的实施比较直观：如果构造的马氏链在时刻 t 处于状态 x，即 $X_t = x$，则首先由 $q(\cdot|x)$ 产生一个潜在的转移 $x \to y$，然后以 $\alpha(x, y)$ 的概率接受 y 作为马氏链的下一时刻的状态值，即 $X_{t+1} = y$；而以概率 $1-\alpha(x, y)$ 拒绝转移到 y，从而马氏链在下一时刻仍处于状态 x，$X_{t+1} = x$。

如果选取

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} \tag{13.6.6}$$

此时有

$$p(x, y) = \begin{cases} q(x, y), & \pi(y)q(y, x) \geqslant \pi(x)q(x, y) \\ q(y, x)\dfrac{\pi(y)}{\pi(x)}, & \pi(y)q(y, x) < \pi(x)q(x, y) \end{cases}$$

可证明由上述过程产生的马氏链是时间可逆的，即 $\pi(x)p(x, y) = \pi(y)p(y, x)$，且 $\pi(x)$ 是马氏链的平稳分布。

下面总结上述用于生成 $\pi(x)$ 为平稳分布的时间可逆马氏链的 Metropolis - Hastings 算法。

（1） 选择一个转移概率为 $q(x, y)$，$x, y = 1, \cdots, N$ 的不可约马氏转移矩阵，并选择某个 1 到 N 的整数 x；

（2） 令 $n=0$，$X_0 = x$；

（3） 生成一个随机变量 X 的随机数使得 $P(X_n = y) = q(X_n, y)$；

---

[1] Aarts and Korst, Simulated Annealing and Bolltzmann Machines, Wiley, New Yok, 1989.

（4）生成一个在 0 到 1 间均匀分布的随机数 U，若 U 小于等于 $\alpha(x, y)$，则 $NS=X$，若 U 大于 $\alpha(x, y)$，则 $NS=X_n$；

（5）$n=n+1$，$X_n=NS$

（6）转到第（3）步。

13.6.4 吉布斯抽样

应用最广泛的 Metropolis - Hastings 方法是吉布斯抽样。设随机向量 $X=(X_1, \cdots, X_n)$ 的概率分布函数为 $p(x)$，其中 $p(x)$ 仅需确定一个乘积常数，即

$$p(x)=cg(x) \tag{13.6.7}$$

其中 $g(x)$ 已知，而 c 未知。

对状态 $x=(x_1, \cdots, x_n)$ 的马氏链应用 Metropolis - Hastings 算法，其转移概率定义如下：如果当前状态为 $\mathbf{x}$，从 $1, \cdots, n$ 中等可能的选取一个坐标。如果选取坐标 i，则可生成一个随机变量 X，其概率分布函数为分量 X_i 关于其他分量的条件分布，即

$$P(X=x)=P(X_i=x \mid X_j=x_j, j \neq i) \tag{13.6.8}$$

且如果 $X=x$，则 $\mathbf{y}=(x_1, \cdots, x, x_{i+1}, \cdots, x_n)$ 为下一个坐标。换句话说，已知 $\mathbf{x}$ 和 $\mathbf{y}$，吉布斯抽样利用 Metropolis-Hastings 算法，有

$$q(\mathbf{x}, \mathbf{y})=\frac{1}{n} P(X_i=x \mid X_j=x_j, j \neq i)=\frac{p(y)}{nP(X_j=x_j, j \neq i)}$$

由于极限分布为 p，根据式（13.6.6），向量 $\mathbf{y}$ 作为新状态的概率为：

$$\alpha(\mathbf{x}, \mathbf{y})=\min\left(\frac{p(\mathbf{y}) q(\mathbf{y}, \mathbf{x})}{p(\mathbf{x}) q(\mathbf{x}, \mathbf{y})}, 1\right)=\min\left(\frac{p(\mathbf{y}) p(\mathbf{x})}{p(\mathbf{x}) p(\mathbf{y})}, 1\right)=1$$

因此，利用吉布斯抽样得到的坐标总可以作为马氏链的下一步状态。

吉布斯抽样具体步骤可如下进行：在给出起始点 $\mathbf{x}^{(0)}=(x_1^{(0)}, \cdots, x_n^{(0)})$，假定已知第 $t+1$ 次抽样开始时的观测值为 $\mathbf{x}^{(t)}=(x_1^{(t)}, \cdots, x_n^{(t)})$，则第 $t+1$ 次抽样的具体步骤如下：

- 由条件分布 $\pi(x_1 \mid x_2^{(t)}, \cdots, x_n^{(t)})$ 抽取 $x_1^{(t+1)}$；

  ……

- 由条件分布 $\pi(x_i \mid x_1^{(t+1)}, \cdots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \cdots, x_n^{(t)})$ 抽取 $x_i^{(t+1)}$；

  ……

- 由后验分布 $\pi(x_n \mid x_1^{(t+1)}, \cdots, x_{n-1}^{(t+1)})$ 抽取 $x_n^{(t+1)}$。

记 $\mathbf{x}^{(t+1)}=(\mathbf{x}_1^{(t+1)}, \cdots, \mathbf{x}_n^{(t+1)})$，最终得到 $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(t)}, \mathbf{x}^{(t+1)}, \cdots$

由 $q(x, y)$ 的构造知，Q 是一个不可约的转移矩阵。因此马氏链是遍历性的。由不同 $x^{(0)}$ 出发，当 $t \to \infty$ 时，可以认为各时刻 $\mathbf{x}^{(t)}$ 的边际分布为极限分布，此时它收敛。而在收敛出现前的 m 次迭代中，各状态的边际分布还不能认为是 $\pi(x)$，因此在估计 $E_\pi(h(X))$ 时应将前 m 个迭代值去掉，即：

$$\frac{1}{n-m}\sum_{i=1}^{n-m} h(x_i) \to E_\pi(h(X))$$

在贝叶斯理论中，未知参数的后验分布大多是多维的、复杂的分布函数，计算与后验分布有关的概率或函数期望往往很困难，这时可应用吉布斯抽样来解决。假设 $X_1, X_2, \cdots, X_n$ 独立，分布参数为 θ，θ 未知。给定 $\Theta=\theta$ 时，X_j 的分布为 $f_{X_j|\Theta}(x_j|\theta)$，则 $\vec{X}=(X_1, X_2, \cdots, X_n)$ 的条件分布密度为 $\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)$。设 Θ 的分布密度为 $\pi(\Theta)$，在贝叶斯估计中，在已知 $\vec{X}=\vec{x}$ 条件下，$\Theta=(\Theta_1, \cdots, \Theta_n)$ 的后验分布为：

$$\pi_{\Theta|\vec{X}}(\theta|\vec{x}) = \frac{f_{\vec{X},\Theta}(\vec{x}, \theta)}{f_{\vec{X}}(\vec{x})} = \frac{1}{f_{\vec{X}}(\vec{x})}\left(\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\pi(\theta)\right) \tag{13.6.9}$$

显然，后验分布满足吉布斯抽样所要求的分布形式（13.6.7），因此可用吉布斯抽样来解决有关后验分布的计算问题。下面是一个具体的例子。

【例 13－18】① 设 $N_1(t)$ 为一个棒球赛季前 $100t\%$ 的赛程中球员 A 本垒打的次数，$0\leqslant t\leqslant 1$。类似地，设 $N_2(t)$ 为球员 B 的本垒打次数。

设存在随机变量 W_1 和 W_2，满足在给定 $W_1=w_1$，$W_2=w_2$ 下，$\{N_1(t), 0\leqslant t\leqslant 1\}$ 和 $\{N_2(t), 0\leqslant t\leqslant 1\}$ 为独立泊松过程，且参数分别为 w_1 和 w_2。此外，设 W_1 和 W_2 是参数为 Y 的独立指数随机变量，且 Y 自身为（0.02，0.10）上均匀分布的随机变量。换言之，假设球员击出本垒打的次数服从一个参数为随机变量的泊松过程，且这一随机变量的分布的参数本身服从已知分布。

假设在前半赛季 A 击出 25 次本垒打，B 击出 18 次。根据吉布斯抽样来估计他们个人在全赛季分别击出的本垒打均值。

解：由题知，存在随机变量 Y、W_1 和 W_2 满足：

（1）Y 服从（0.02，0.10）上均匀分布；

（2）给定 $Y=y$，W_1 和 W_2 是参数为 y 的独立同分布的指数随机变量；

（3）给定 $W_1=w_1$，$W_2=w_2$，$\{N_1(t), 0\leqslant t\leqslant 1\}$ 和 $\{N_2(t), 0\leqslant t\leqslant 1\}$ 为独立泊松过程，且参数分别为 w_1 和 w_2。

为求 $E[N_1(1)|N_1(0.5)=25, N_2(0.5)=18]$，首先加上 W_1 的条件：

$$E[N_1(1)|N_1(0.5)=25, N_2(0.5)=18, W_1]=25+0.5W_1$$

求在给定 $N_1(0.5)=25$ 和 $N_2(0.5)=18$ 下的条件概率，得：

$$E[N_1(1)|N_1(0.5)=25, N_2(0.5)=18]$$
$$=25+0.5E[W_1|N_1(0.5)=25, N_2(0.5)=18]$$

① 本节两个例子来自罗斯（Sheldo M. Ross）著、王兆军等译：《统计模拟》（第 4 版），人民邮电出版社 2007 年版，第 214 页。

类似地，有

$$E[N_2(1)\mid N_1(0.5)=25,\ N_2(0.5)=18]$$
$$=18+0.5E[W_2\mid N_1(0.5)=25,\ N_2(0.5)=18]$$

现利用吉布斯抽样估计这些条件期望。首先，对于 $0.02<y<0.10$，$w_1>0$，$w_2>0$，有

$$f(y,\ w_1,\ w_2,\ N_1(0.5)=25,\ N_2(0.5)=18)$$
$$=Cy^2e^{-(w_1+w_2)y}e^{-(w_1+w_2)/2}(w_1)^{25}(w_2)^{18}$$

其中 C 不依赖于 y，w_1，w_2。因此，对于 $0.02<y<0.10$，有

$$f(y\mid w_1,\ w_2,\ N_1(0.5)=25,\ N_2(0.5)=18)=C_1y^2e^{-(w_1+w_2)y}$$

上式说明给定 w_1，w_2，$N_1(0.5)=25$，$N_2(0.5)=18$ 下 Y 的条件分布为参数为 3 和 w_1+w_2 的伽玛随机变量在 0.02 到 0.10 间取值的条件分布。另外，

$$f(w_1\mid y,\ w_2,\ N_1(0.5)=25,\ N_2(0.5)=18)=C_2e^{-(y+1/2)w_1}(w_1)^{25}$$

由此可知，在给定 y，w_2，$N_1(0.5)=25$，$N_2(0.5)=18$ 下 W_1 的条件分布是参数为 26 和 $y+0.5$ 的伽玛分布。同理，在给定 y，w_1，$N_1(0.5)=25$，$N_2(0.5)=18$ 下 W_2 的条件分布是参数为 19 和 $y+0.5$ 的伽玛分布。

于是，从 y，w_1，w_2 的值开始，其中 $0.02<y<0.10$，且 $w_i>0$，吉布斯抽样如下进行：

（1）产生一个值在 0.02 与 0.10 之间的参数为 3 和 w_1+w_2 的伽玛随机变量的值，并令其为新的 y 值；

（2）产生参数为 26 和 $y+0.5$ 的一个伽玛随机变量的值，并令其为新的 w_1；

（3）产生参数为 19 和 $y+0.5$ 的一个伽玛随机变量的值，并令其为新的 w_2；

（4）回到步骤（1）。

w_1 的平均值即为 $E[W_1\mid N_1(0.5)=25,\ N_2(0.5)=18]$ 的估计，w_2 的平均值即为 $E[W_2\mid N_1(0.5)=25,\ N_2(0.5)=18]$ 的估计。前者的一半加上 25 即为我们对 A 全年击出的本垒打的平均数的估计，后者的一半加上 18 为 B 击出的本垒打的平均数估计。 ■

【例 13-19】 设 X_i，$i=1, 2, 3, 4, 5$ 为独立指数随机变量，且 X_i 的均值为 i，使用吉布斯抽样利用模拟方法估计

$$\beta=P\{\prod_{i=1}^{5}X_i>120\mid\sum_{i=1}^{5}X_i=15\}$$

解： 随机选择坐标中的两个就可以完成此项任务。首先，假设 X 和 Y 是参数分别为 λ 和 μ 的独立指数随机变量，其中 $\mu<\lambda$，并且我们按如下方法寻找在 $X+Y=a$ 下的 X 的条件分布：

$$f_{X\mid X+Y}(x\mid a)=C_1f_{X,Y}(x,\ a-x),\ 0<x<a$$

$$= C_2 e^{-\lambda x} e^{-\mu(a-x)}, \ 0 < x < a$$
$$= C_3 e^{-(\lambda-\mu)x}, \ 0 < x < a$$

此条件分布为一小于 a 条件下的参数 $\lambda-\mu$ 的指数随机变量的条件分布。

利用这一结果，通过令初始状态$(x_1, x_2, x_3, x_4, x_5)$是和等于15的任意5个正数，我们可以估计$\beta$。现从1，2，3，4，5中随机选出两个数；假设挑选的两个数为 $I=2$ 和 $J=5$，则在给定其他值下 X_2 和 X_5 的条件分布为在其和等于 $15-x_1-x_3-x_4$ 下两个均值分别为2和5的独立指数随机变量的条件分布。但是，由前所述，要获得 X_2 和 X_5 的值，首先要产生一个小于 $15-x_1-x_3-x_4$ 且参数为 $0.5-0.2=0.3$ 的指数随机变量的值。之后，令 x_2 等于这个值，且重设 x_5 使得 $\sum_{i=1}^{5} x_i = 15$。不断重复这一过程，用 $\prod_{i=1}^{5} X_i > 120$ 的状态向量 x 的比例作为 β 的估计。■

§13.7　精算建模中的随机模拟实例

本节以例题的方式给出一些模拟在精算模型中的应用，这些内容只是在前面模拟的基本方法中加入保险的背景。

【例 13－20】 假设索赔次数服从泊松分布，每年索赔次数的平均值为1。索赔额服从均值为10 000，标准差为500的正态分布。根据[0，1]区间上均匀分布的随机数列0.4、0.8模拟前两年的索赔次数；根据[0，1]区间上均匀分布的随机数列0.1、0.3和0.5模拟每次索赔额。保险公司对于每年的索赔额的免赔额为5 000。计算这两年内保险公司给付总数的一个模拟结果。

解： 由于索赔次数服从均值为1的泊松分布，分布函数如表13－7所示。

表 13－7

N	0	1	2	3
$f(n)$	0.368	0.368	0.184	0.061
$F(n)$	0.368	0.736	0.92	0.981

若用随机数0.4和0.8进行模拟，因为

$$F(0)=0.368 \leqslant 0.4 < F(1)=0.736, \ F(1)=0.736 \leqslant 0.8 < F(2)=0.92$$

根据反函数原理可得第一年索赔次数为1，第二年的索赔次数为2，两年内索赔次数的模拟值是3，所以要模拟3个索赔额。

对于索赔额根据反函数法，根据[0，1]区间上均匀分布的随机数列0.1、0.3和0.5可以得到标准正态分布的随机数 Z 为 -1.282，-0.2544，0。

损失 X 服从均值为10 000，标准差为500的正态分布，所以令 $X=10\,000+500Z$，可得模拟的损失随机数为9 359，9 872.8，10 000。

根据保险公司对于每年索赔额的免赔额为5 000，则第一年保险公司的损失为9 359，赔付为4 359，第二年保险公司的损失为19 872.8，赔付为

14 872.8。所以，保险公司这两年的总赔付为 19 231.8。 ■

【例 13 -21】 假设索赔次数服从二项分布 $B(4, 0.5)$，索赔强度服从 $\alpha=2$，$\theta=1\ 000$ 的帕累托分布。用均匀分布随机数 0.81，0.53，0.68，0.12 来模拟 N，X_1，X_2，…，赔偿限额为 1 000，根据随机数得到模拟值计算保险在第几起损失发生时完全不负赔付责任。

解： 表 13 -8 是索赔次数 N 对应的分布列。因此均匀随机数 0.81 对应的索赔次数为 3。

表 13 -8

N	0	1	2	3	4
$f(n)$	0.0625	0.25	0.375	0.25	0.0625
$F(n)$	0.0625	0.3125	0.6875	0.9375	1

帕累托分布的分布函数 $F(x)=1-\left(\frac{\theta}{\theta+x}\right)^{\alpha}=1-\left(\frac{1\ 000}{1\ 000+x}\right)^{2}$，所以 $x=1\ 000\sqrt{\frac{1}{1-u}}-1\ 000$。那么 0.53，0.68，0.12 对应的赔付额为 458.65，767.77，66.0。因为 $458.65<1\ 000<458.65+767.77$，所以保险公司对第二起损失部分赔偿，对第三起损失完全不负赔付责任。 ■

【例 13 -22】 假设索赔次数服从 Possion（3），理赔额服从均值为 1 000，标准差为 600 的正态分布。假设初始盈余为 1 000，安全附加系数 θ 为 0.1。保费收取在年初，当盈余为负时保险公司则会破产。从随机数表选出的在［0，1］区间内的随机数 0.23，0.14，0.49，0.34，0.21 来模拟理赔的时间间隔，用随机数 0.50，0.17，0.88，0.62，0.74 来模拟赔付额。求该保险公司在何时破产。

解： 由题目得：

$$期望保费=期望索赔次数\times期望赔付额\times(1+\theta)$$

所以，期望保费 $=3\times1\ 000\times1.1=3\ 300$。

由泊松分布的性质，保险公司发生赔付的时间间隔服从均值为 1/3 的指数分布。设模拟得到的时间间隔为 t，则有

$$r_i=1-e^{-3t_i},\qquad t_i=-\frac{1}{3}\ln(1-r_i)$$

则模拟得到的 t 依次为 0.087，0.0502，0.2244，0.1385，0.0786。

对于索赔额，根据反函数法，根据［0，1］区间上均匀分布 R 的随机数列 0.50，0.17，0.88，0.62，0.74 可以得到标准正态分布的随机数 Z 为 0，-0.9542，1.1750，0.3055，0.6433。

损失 X 服从均值为 1 000，标准差为 200 的正态分布，所以令 $X=1\ 000+$

$200Z$，可得模拟的损失随机数为 1 000，809.16，1 235，1 061.1，1 128.66。

下面我们开始模拟保险公司的损益过程：

- 在时刻 0，保险公司的盈余 = 初始盈余 + 收到的保费 = 3 300 + 1 000 = 4 300；
- 在时刻 0.087，发生第一次索赔，保险公司此刻的盈余 = 4 300 − 1 000 = 3 300；
- 在时刻 0.1372，发生第二次索赔，保险公司此刻的盈余 = 3 300 − 809.16 = 2 490.84；
- 在时刻 0.3616，发生第三次索赔，保险公司此刻的盈余 = 2 490.84 − 1 235 = 1 255.84；
- 在时刻 0.5001，发生第四次索赔，保险公司此刻的盈余 = 1 255.84 − 1 061.1 = 194.74；
- 在时刻 0.5787，发生第五次索赔，保险公司此刻的盈余 = 194.74 − 1 128.66 < 0。

所以保险公司在时刻 0.5787 破产。 ■

在这个模型中可以看出，保险公司每次的索赔数额的离差不大，但是索赔次数一旦发生多次，保险公司便难以承受。可以建议保险公司提高安全附加系数或者购买责任超赔再保险来保证公司的正常运营。

【例 13－23】 假设索赔的时间间隔服从均值为 1/3 的指数分布，理赔额 $X=10^t$，t 为索赔发生的时刻。保费的收取为一个连续的过程，收取率为 $20t^4$ 且 $i=0$。当盈余为负时保险公司则会破产。从随机数表选出的在 (0，1) 区间内的随机数 0.5，08，0.9 来模拟理赔的时间间隔，求在已知情况下的初始的最小资金需求来保证在这三次索赔下保险公司的正常运营。 ■

解：设模拟得到的时间间隔为 m，则有

$$r_i = 1 - e^{-3m_i},\quad m_i = -\frac{1}{3}\ln(1-r_i)$$

则模拟得到的 m 依次为 0.231、0.5365、0.7675。

根据理赔额 $X=10^t$，t 为索赔发生的时刻，模拟得到的 t 依次为 0.231、0.7675、1.535，则依次得到的 X 依次为 1.7022、5.8546、34.2768。

根据保费的收取率 $20t^4$，在时刻 0.231、0.7675、1.535，收到的保费为 $\int_0^{t_i} 20t^4 dt$，根据公式可得到在时刻 0.231、0.7675、1.535 收取的保费依次为 0.0026、1.06、34。

设初始的最小资本需求为 c，那么，

- 在时刻 0.231，为了保证保险公司正常运营，$c+0.0026-1.7022>0$；
- 在时刻 0.5365，为了保证保险公司的正常运营，$c+1.06-1.7022$

$-5.8546>0$;

- 在时刻 1.535，为了保证保险公司的正常运营，$c+34-1.7022-5.8546-34.2768>0$。

可以得出 $c>7.8336$，即必须保证初始的最小资金大于 7.8336 才能使保险在这三次索赔发生后能仍能正常运营。 ■

习　题

1. 根据[0, 1]区间上均匀分布的随机数列 0.3、0.6875 和 0.95 表示二项分布 $B(4, 0.5)$ 的数。

2. 根据[0, 1]区间上均匀分布的随机数列 0.1247、0.9321 和 0.6873 来表示 Possion (3) 的数。

3. 一个科学家做实验，成功率为 0.6，X 表示到第一次成功的试验次数。根据[0, 1]区间上均匀分布 R 的随机数列 0.85、0.38、0.63、0.22 来模拟 X。计算到第三次成功的试验次数。

4. 假设

$$F_X(x)=\begin{cases}0.5x, & 0\leqslant x<1\\ 0.5+0.25x, & 1\leqslant x\leqslant 2\end{cases}$$

根据[0, 1]区间上均匀分布的随机数列 0.3、0.6 和 0.9 模拟 X 的随机数列。

5. 随机变量 X 的分布函数 $F_X(x)$ 是两个指数分布的综合，分布 1 是均值为 1 的指数分布，权重为 0.25；分布 2 是均值为 2 的指数分布，权重为 0.75。根据[0, 1]区间上均匀分布的随机数 0.7 来模拟 X。

6. 假设随机变量 X 的样本均值为 $\bar{X}$，方差为 s^2。通过模拟来估计总体均值，求使得真值与估计值的相对误差不大于 5% 的概率为 0.9 所需的模拟次数。

7. 假设通过模拟得到 $X_1=1$，$X_2=2$，$X_3=3$，$X_4=4$，$X_5=5$，为了使 $E(X)$ 估计值的标准差不大于 0.05 所需的模拟次数。

8. 设随机变量 X 服从指数分布。通过模拟来估计总体 $F_X(100)$。假设 P_n 为样本中小于 100 的数目，n 为样本数。求使得真值与估计值的相对误差不大于 5% 的概率为 0.9 所需的模拟次数。

9. 存在一个随机样本，样本的分布函数未知，已知样本标准差的区间为[2, 3]，求使得样本均值的 0.9 置信区间不大于 1 的最小样本量。

10. 假设某个理赔员处理一次索赔时间为 0.5 个小时或 1 小时，概率分别为 0.5，小的随机数对应小的处理时间，随机数为 0.1，0.6，0.4；用均

匀分布随机数0.2、0.4、1.1来表示索赔事件在某2个小时时间段内发生的时间。问该理赔员在该时段结束时处理索赔的状态。

11. 假设一个健康险的分布为符合泊松分布，索赔次数服从Possion(3)，每次索赔额服从的分布函数为 $F_X(x)=\begin{cases}0.5x, & 0\leqslant x<1\\ 0.5+0.25x, & 1\leqslant x\leqslant 2\end{cases}$，单位为万元。根据[0，1]区间上均匀分布 R 的随机数列0.1247、0.4121模拟前两年的索赔次数；根据[0，1]区间上均匀分布 R 的随机数列0.3、0.6和0.9模拟每次索赔额。保险公司对于每年的索赔额的免赔额为5 000元。计算这两年内保险公司给付总数的一个模拟结果。

12. 假设索赔次数服从二项分布($n=4$，$p=0.5$)，索赔强度服从均值为1 000的指数分布，用均匀分布随机数0.21，0.53，0.67，0.13来模拟 N，X_1，X_2，…，计算总赔付额。

13. 假设一个车险的每月的损失分布服从均值为1 000的指数分布，每月的免赔额为300，根据[0，1]区间上均匀分布 R 的随机数列0.213，0.376，0.754，0.109模拟前四个月的损失额，计算保险公司前四个月的赔付额。

14. 索赔次数服从二项分布（4，0.5），赔付额服从帕累托分布（2.5，1 000）。根据[0，1]区间上均匀分布 R 的随机数列0.2，0.8，0.3，0.1，0.5，0.6，0.9，0.3来模拟索赔次数和索赔额，当模拟的总索赔次数达到4时停止模拟。计算保险公司的总赔付额。

15. 假设索赔次数服从Possion（3），理赔额服从帕累托分布（2，1 000）。假设初始盈余为1 000，安全附加为0.1，保费收取在年初，当盈余为负时保险公司则会破产。从随机数表选出的在（0，1）区间均匀分布内的随机数0.23，0.94，0.49，0.34，0.21来模拟理赔的时间间隔，用随机数0.58，0.97，0.88，0.67，0.44来模拟赔付额。求该保险公司在何时破产。

16. 假设索赔的时间间隔服从均值为1/3的指数分布，理赔额 $X=10^t$，t 为索赔发生的时刻。假设初始盈余为5，保费的收取为一个连续的过程，收取率为 ct^4 且 $i=0$。当盈余为负时保险公司则会破产。随机产生在（0，1）区间均匀分布随机数0.5，0.8，0.9来模拟理赔的时间间隔，求在已知情况下 c 的条件来保证在这三次索赔下保险公司的正常运营。

17. 随机抽取随机变量 X 的三个样本：1，6，8，应用Bootstrap方法计算下面估计的均方误差。

(1)均值估计；

(2)max(X)；

(3)min(X)。

18. 假设随机变量 X_1、X_2、…、X_{10}独立同分布于 $F(x)$，现有它们的一组观测值：1.2、3.2、5.3、6.4、3.6、3.7、6.0、5.4、3.1、3.9。下面是被模拟的经验分布的自助样本，每个样本有10个数据：

样本1：3.6、3.7、3.7、3.9、5.4、1.2、3.2、3.2、3.9、6.4

样本2：1.2、1.2、5.3、5.3、3.9、3.2、3.2、3.1、5.4、6.4

样本3：3.9、3.9、6.0、6.0、5.4、1.2、3.9、6.0、1.2、1.2

使用这三个自助样本估计样本方差估计值的MSE。

19. 假设随机变量 X_1、X_2、…、X_{10}独立同分布于 $F(x)$，现有它们的一组观测值：1.2、3.2、5.3、6.4、3.6、3.7、6.0、5.4、3.1、3.9。下面是被模拟的经验分布的自助样本，每个样本有10个数据：

样本1：3.6、3.7、3.7、3.9、5.4、1.2、3.2、3.2、3.9、6.42

样本2：1.2、1.2、5.3、5.3、3.9、3.2、3.2、3.1、5.4、6.4

样本3：3.9、3.9、6.0、6.0、5.4、1.2、3.9、6.0、1.2、1.2

使用这三个自助样本估计 $P(X<5)$估计值的MSE。

20. 假设随机变量 X_1、X_2、…、X_{10}独立同分布于 $F(x)$，现有它们的一组观测值：1.2、3.2、5.3、6.4、3.6、3.7、6.0、5.4、3.1、3.9。下面是被模拟的经验分布的自助样本，每个样本有10个数据：

样本1：3.6、3.7、3.7、3.9、5.4、1.2、3.2、3.2、3.9、6.4

样本2：1.2、1.2、5.3、5.3、3.9、3.2、3.2、3.1、5.4、6.4

样本3：3.9、3.9、6.0、6.0、5.4、1.2、3.9、6.0、1.2、1.2

使用这三个自助样本估计0.3分位点估计值的MSE。

第十四章 案例分析

学习目标

□ 了解精算建模的一般过程

§14.1 引 言

精算建模的目的不在于完成一次数学运算，也不在于提供一个关于保险环境的过分复杂的描述。其实，建立并分析模型的目的是回答关于保险赔付、投资组合及管理运营方面的一些重大问题。

我们已经在前面各章讨论了基本风险模型的性质、参数的估计、模型的选择与调整等一系列问题。这些问题都是分开讨论的，本章将用两个案例来具体说明精算建模的过程。第一个例子是关于退休人员的养老金问题；第二个例子是关于再保险定价问题。

值得注意的是，我们所讨论的案例都是对前面介绍的多种方法的综合运用。案例的假设或承保责任可能简化或复杂了实际情况，但可以帮助我们考虑解决各方面的问题。

§14.2 退休人员的死亡时间和养老金

14.2.1 案例介绍

下面我们将用实例说明一个经验生命表的构建过程，并以生命表为基础讨论寿险赔付的一些基本计算。假设我们要研究退休后人群的死亡规律，表 14－1 给出了调查得到的暴露数和实际死亡数[1]。

我们把死亡人数 θ_x 看成是基于样本规模 n_x 的二项分布随机变量，于是 $q_x^0=\theta_x/n_x$ 作为真实死亡率 q_x 的初始估计值（表 14－1 中称之为“粗死亡率”）。图 14－1 描述了不同年龄的粗死亡率值，直观上看这些估计值在高龄部分不是光滑的，有必要进行进一步的建模处理。

---

① 该数据是近似数据，来自于 Dick London 著、徐诚浩译：《修匀数学》，上海科学技术出版社 1996 年版，第 203 页。

表 14－1　　　　　　暴露数和实际死亡数

年龄 x	死亡数 θ_x	暴露数 n_x	粗死亡率 q_x^0	估计标准差	年龄 x	死亡数 θ_x	暴露数 n_x	粗死亡率 q_x^0	估计标准差
55	1	84	0.01190	0.011834	80	374	6 140	0.06091	0.003052
56	2	418	0.00478	0.003375	81	348	4 718	0.07376	0.003805
57	10	1 066	0.00938	0.002953	82	304	3 791	0.08019	0.004411
58	21	2 483	0.00846	0.001838	83	249	2 806	0.08874	0.005368
59	35	3 721	0.00941	0.001582	84	167	2 240	0.07455	0.00555
60	62	5 460	0.01136	0.001434	85	192	1 715	0.11195	0.007614
61	50	6 231	0.00802	0.00113	86	171	1 388	0.12320	0.008822
62	55	8 061	0.00682	0.000917	87	126	898	0.14031	0.01159
63	88	9 487	0.00928	0.000984	88	86	578	0.14879	0.014803
64	132	10 770	0.01226	0.00106	89	97	510	0.19020	0.017378
65	267	24 267	0.01100	0.00067	90	93	430	0.21628	0.019854
66	300	26 791	0.01120	0.000643	91	75	362	0.20718	0.021301
67	432	29 174	0.01481	0.000707	92	84	291	0.28866	0.026564
68	491	28 476	0.01724	0.000771	93	31	232	0.13362	0.022338
69	422	25 840	0.01633	0.000788	94	75	196	0.38265	0.034717
70	475	23 916	0.01986	0.000902	95	29	147	0.19728	0.032822
71	413	21 412	0.01929	0.00094	96	25	100	0.25000	0.043301
72	480	20 116	0.02386	0.001076	97	20	161	0.12422	0.025995
73	537	18 876	0.02845	0.00121	98	5	11	0.45455	0.150131
74	566	17 461	0.03242	0.00134	99	3	10	0.30000	0.144914
75	581	15 012	0.03870	0.001574	100	2	8	0.25000	0.153093
76	464	11 871	0.03909	0.001779	101	0	5	0.00000	0
77	461	10 002	0.04609	0.002097	102	2	4	0.50000	0.25
78	433	8 949	0.04839	0.002268	103	0	2	0.00000	0
79	515	7 751	0.06644	0.002829	104	1	2	0.50000	0.353553

关于死亡率序列的真实模式应满足以下三个要素：

（1）它是光滑的；

（2）它是随 x 而递增的；

（3）在高年龄区间内，曲线呈更陡的上升趋势。

在图 14－1 中，q_x^0 的基本趋势满足先验观点要素（2）和（3）。但是在 55～65 岁的死亡率具有小幅波动，特别是 55 岁和 60 岁的死亡率明显高于其他相近年龄；另外在 91 岁以后的死亡率呈剧烈的锯齿状，101 岁和 103 岁的死亡率等于 0。这与先验观点不符。从样本数可以看出，55 岁、

98～104 岁的样本数较少，估计标准差相应也比较大，因此这些估计值的可信度不高。修匀的目的就是要改进这些初始估计值，使其满足先验的观点。

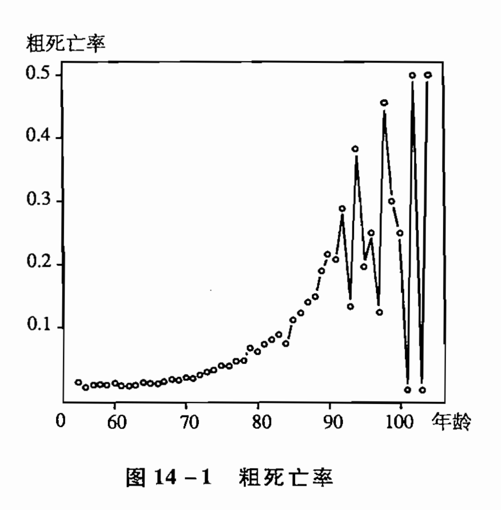


图 14－1 粗死亡率

在这里，我们以 Whittaker 修匀和样条修匀为例说明表格和参数修匀的过程，对每一种情形，都给出方法的叙述，且用表格和图形两种形式介绍所得的结果。用 $w_x=\dfrac{n_x}{v_x(1-v_x)}$ 计算每个结果的拟合度量，光滑度量为 $S=\sum\limits_x(\Delta^3v_x)^2$。对每种单个修匀方法都给出这两种度量。在图上也可观察拟合性和光滑性。

14.2.2 死亡率的修匀

1. Whittaker 修匀。用式（11.2.10）给出的基本方法施行以下两种修匀方法，在两种情形下都取 $z=4$，所计算的权重为：

$$w_x=\frac{n_x}{q_x^0(1-q_x^0)}$$

为了说明参数 h 的作用，我们取两个不同的 h 值。在图 14－2 中分别取 $h=1\ 000$ 和 $h=426\ 138$（w_x 的平均值）。正如 11.2.2 中所讨论的那样，在修匀结果中，大的 h 值产生大的光滑性，但是却偏离了初始估计。所得结果在表 14－2 中表示。

为了避免在计算权时出现分母是 0 的情形（初始估计值 $q_{101}^0=q_{103}^0=0$），可任意用 $q_{101}^0=0.25$ 和 $q_{103}^0=0.5$ 代替之。

表 14－2 修匀结果

年龄 x	Whittaker 修匀值		样条修匀值	年龄 x	Whittaker 修匀值		样条修匀值
	$h=1\ 000$	$h=426\ 138$			$h=1\ 000$	$h=426\ 138$	
55	0.00990	0.00387	0.00715	61	0.00810	0.00830	0.00878
56	0.00548	0.00620	0.00747	62	0.00681	0.00813	0.00920
57	0.00854	0.00815	0.00773	63	0.00931	0.00902	0.00973
58	0.00862	0.00939	0.00796	64	0.01218	0.01016	0.001041
59	0.00954	0.00971	0.00819	65	0.01102	0.01102	0.01126
60	0.01115	0.00916	0.00846	66	0.01121	0.01226	0.01232

续表

年龄 x	Whittaker 修匀值		样条修匀值	年龄 x	Whittaker 修匀值		样条修匀值
	$h=1\ 000$	$h=426\ 138$			$h=1\ 000$	$h=426\ 138$	
67	0.01481	0.01430	0.01361	86	0.12349	0.11976	0.12119
68	0.01718	0.01624	0.001516	87	0.14048	0.13935	0.13325
69	0.01645	0.001746	0.01701	88	0.15951	0.16085	0.14615
70	0.01969	0.01862	0.01918	89	0.18469	0.18209	0.15959
71	0.01943	0.02047	0.02170	90	0.20726	0.20069	0.17298
72	0.02378	0.2372	0.02459	91	0.21917	0.21453	0.18572
73	0.02840	0.02803	0.02790	92	0.22253	0.22212	0.19722
74	0.03263	0.03257	0.03164	93	0.22449	0.22290	0.20686
75	0.03828	0.03683	0.03585	94	0.23026	0.21739	0.21405
76	0.03972	0.04102	0.04055	95	0.22027	0.20708	0.21819
77	0.04493	0.04593	0.04578	96	0.19142	0.19453	0.21867
78	0.05036	0.05197	0.05156	97	0.16012	0.18312	0.21488
79	0.06290	0.05877	0.05792	98	0.15134	0.17683	0.20624
80	0.06387	0.06540	0.06489	99	0.17660	0.17992	0.19213
81	0.07225	0.07140	0.07251	100	0.23486	0.19680	0.17196
82	0.08109	0.07698	0.08079	101	0.31745	0.23194	0.14511
83	0.08313	0.08317	0.08977	102	0.41135	0.28982	0.11100
84	0.08352	0.09157	0.09948	103	0.50127	0.37493	0.06901
85	0.10236	0.10362	0.10994	104	0.57134	0.49176	0.01885

两种结果的拟合性与光滑性度量分别为：

$h=1\ 000$：　$F=68.12450$，　$S=0.00461$

$h=426\ 138$：　$F=113.40598$，　$S=0.00019$

2. 样条修匀。由于样条修匀是一种参数修匀，本身就是光滑的，所以此时不必计算光滑度量。这里采用两弧三次样条：

$$v_x=\begin{cases}p_0(x), & a\leqslant x\leqslant k\\ p_1(x), & k\leqslant x\leqslant b\end{cases}$$

由于初始估计值在90岁后发生较大波动，因此取内结点 $k=87.5$。得到结果列在表14－2中的“样条修匀值”列。拟合度量值等于 $F=153.78$。

虽然图 14－3 中的样条修匀值是光滑的，但是相比较而言，拟合性度量值比 Whittaker 要高，而且在高年龄段的死亡率没有体现出递增性。因此，如果单调性是重要的，那么 Whittaker 修匀的结果要比样条结果好。在实际应用中，生命表编制不一定只用一种修匀，可以在不同的年龄段采用不同的修匀方法。

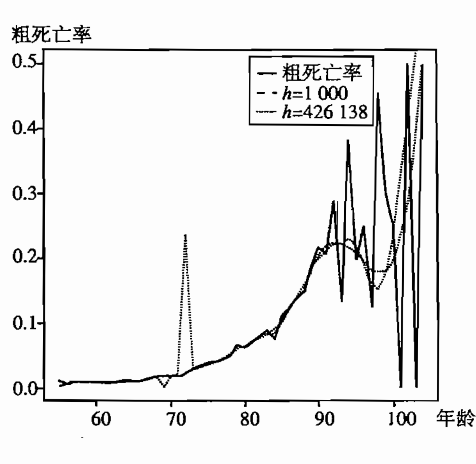


图 14－2　Whittaker 修匀结果

图 14－3　样条修匀值

14.2.3　死亡时间

综合比较修匀结果，Whittaker 修匀结果的拟合性较好，光滑性度量值与样条修匀差别不大。因此我们选定 Whittaker 修匀（$h=1\ 000$）后的死亡率，由此得到 55 岁生存的人在 55 岁以后的生存概率，如表14－3 和图 14－4所示。

表 14－3　　　55 岁以后退休人员的生存函数

年龄 x	死亡率	生存概率$\frac{S(x)}{S(55)}$	年龄 x	死亡率	生存概率$\frac{S(x)}{S(55)}$
55	0.00990	1	65	0.01102	0.913884
56	0.00548	0.9901	66	0.01121	0.903813
57	0.00854	0.984674	67	0.01481	0.893681
58	0.00862	0.976265	68	0.01718	0.880446
59	0.00954	0.96785	69	0.01645	0.86532
60	0.01115	0.958616	70	0.01969	0.851085
61	0.00810	0.947928	71	0.01943	0.834327
62	0.00681	0.94025	72	0.02378	0.818116
63	0.00931	0.933847	73	0.02840	0.798662
64	0.01218	0.925152	74	0.03263	0.77598

续表

年龄 x	死亡率	生存概率$\frac{S(x)}{S(55)}$	年龄 x	死亡率	生存概率$\frac{S(x)}{S(55)}$
75	0.03828	0.750659	90	0.20726	0.18311
76	0.03972	0.721924	91	0.21917	0.145158
77	0.04493	0.693249	92	0.22253	0.113344
78	0.05036	0.662102	93	0.22449	0.088121
79	0.06290	0.628758	94	0.23026	0.068339
80	0.06387	0.589209	95	0.22027	0.052603
81	0.07225	0.551577	96	0.19142	0.041016
82	0.08109	0.511725	97	0.16012	0.033165
83	0.08313	0.470229	98	0.15134	0.027855
84	0.08352	0.431139	99	0.17660	0.023639
85	0.10236	0.39513	100	0.23486	0.019464
86	0.12349	0.354685	101	0.31745	0.014893
87	0.14048	0.310885	102	0.41135	0.010165
88	0.15951	0.267212	103	0.50127	0.005984
89	0.18469	0.224589	104	0.57134	0.002984

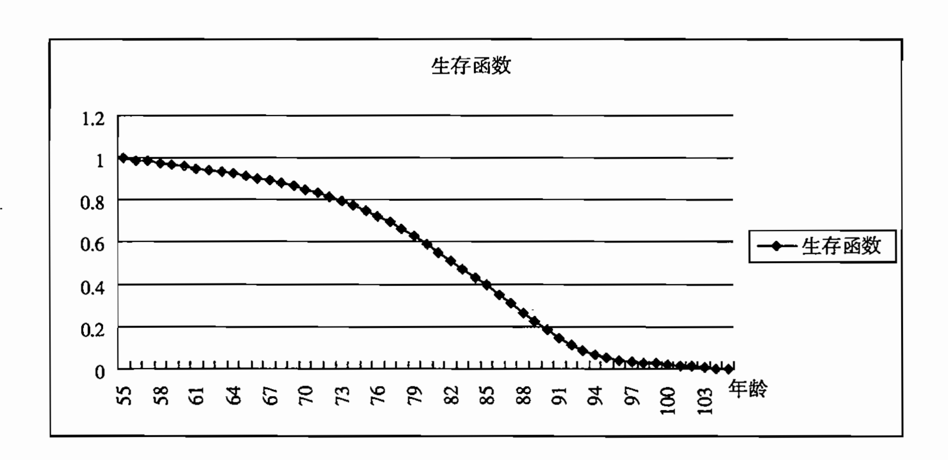


图 14－4　退休人员的生存函数

在生命表 14－3 的基础上，可计算人寿保险的赔付问题。在人寿保险中，人们并不特别关心免赔、限额和共保等条款。我们考虑如下两个问题：

（1）当前为 55 岁退休的个体，退休后在生存期间每年年初领取 1 万元养老金，以 5% 的年利率计算这些现金流的期望现值。

（2）当前为 60 岁的个体，若在死亡时刻给付 1 万元，以 5% 的年利率计算该现金流的期望现值。

对于第一个问题，现值随机变量可以表示为 $Y = 10\ 000(Y_0 + \cdots + Y_{50})$，其中 Y_j 表示生存到 $55+j$ 的条件下受益 1 元的现值，$j = 1, \cdots, 50$。则有

$$Y_j = \begin{cases} 1.05^{-j}, & \dfrac{S(55+j)}{S(55)} \\ 0, & 1 - \dfrac{S(55+j)}{S(55)} \end{cases}$$

结果为：

$$E(Y) = 10\ 000 \sum_{j=0}^{51} \frac{1.05^{-j} S(55+j)}{S(55)} = 144\ 508.56$$

对于第二个问题，先把年利率 5% 转化为连续利率 $\delta = 4.879\%$，令 $Z = 10\ 000 e^{-0.04879T}$ 表示现值随机变量，其中 T 表示 60 岁的个体的未来生存时间，以年为单位。现值期望的计算公式为：

$$E(Z) = 10\ 000 \int_0^{45} e^{-0.04879t} \frac{f(60+t)}{S(60)} dt$$

其中生存函数在整数年龄之间的值由线性插值得到，密度函数为曲线的斜率。将积分区域分为 46 个区间，有

$$\begin{aligned} E(Z) &= 10\ 000 \sum_{j=0}^{45} \frac{(S(60+j) - S(60))/S(55)}{S(60)/S(55)} \int_j^{j+1} e^{-0.04879t} dt \\ &= 10\ 000 \sum_{j=0}^{45} \frac{(S(60+j) - S(60))/S(55)}{S(60)/S(55)} \left(\frac{e^{-0.04879j} - e^{-0.04879(j+1)}}{0.04879} \right) \\ &= 4\ 685.4 \end{aligned}$$

§14.3　再保险定价案例分析

14.3.1　案例介绍

假设你是 FW 再保险公司的定价精算师。FW 公司在全球再保险市场中具有一定地位，并且在承保绝大多数风险方面享有盛誉。现在 FW 公司要求你为一个新的险种——金融董事及高级职员责任保险，做限额损失再保险定价和偿付能力方面的分析。董事及高级职员责任保险承保被保险董事及高级职员在执行职务过程中，由于单独或共同的过错行为导致第三者遭受经济损失，依法应由被保险董事及高级职员承担的赔偿责任。这种责任险保单是损失发生制保单，即保险公司承担上述赔偿责任时以被保险董事及高级职员引起索赔的过错行为发生于保单约定的溯及日后，并且第三者在保险期限内首次向被保险董事及高级职员提出索赔为限。

关于这个险种几乎没有历史数据，所以也没有类似的行业定价数据。不过幸好，FW 公司的这个客户曾经承保过这种业务，拥有关于这方面的赔

付数据。对于每次索赔事件，你的客户提供的数据都含有五个方面的信息：（1）索赔发生的保单年；（2）董事和高级职员的工作领域，这里主要有三类：保险公司、银行以及证券公司；（3）保单承保时的免赔额，包括没有免赔额情形；（4）保单的限额，这里用最大赔付额来表示；（5）保险公司的赔付额。数据见表 14－4 和表 14－5①。

对于责任险来说，当一次索赔发生时，可能需要经过几年甚至几十年，最终的责任额才能确定。因此为了消除时间的影响，这里我们假设表中的数据都是经过趋势化预测的最终损失。

此外，你的客户还提供了损失次数数据。在每个保单年，你都将获得签单的保单数，其中包括三个不同工作类型的董事和高级职员责任险在各保单年的签单数。作为保险合约的一部分，被承保的董事和高级职员被要求提供所有的发生的责任险事故，无论事故的损失额是否超过免赔额。因此，除了前面的索赔额数据外，你还获得了每年的损失次数数据。

作为定价精算师，公司要求你对这个再保险新险种进行定价。根据客户提供的数据，你需要完成下面的任务：

第一，假设三个不同工作领域的董事和高级职员的风险水平是一样的，也就是说，他们发生索赔的可能性一样，索赔额的分布也相同。在这样的假定下，如果对单张保单承保限额损失再保险，即如果保单的原始索赔额低于自留额，则 FW 公司不承担赔付责任；如果高于自留额，则只赔付超过自留额以上的部分原索赔额。除了原保单的保单限额外，我们不考虑再保险限额存在的情况。你被要求估计再保险赔付额的分布，并获得该分布的均值、标准差和 90% 以及 99% 分位数。在此基础上，当保单组合为 100 张保单，且每张保单都具有相同的自留额时，计算再保险总赔付额的分布。

第二，考虑总损失限额再保险情形。总损失限额再保险是对一个日历年内签单的所有保单总赔付额进行再保险，当总赔付额超过总自留额时，FW 公司将对超过的部分进行赔付。假设你已经知道这些保单的数目，如 100 份保单。假设这 100 份保单具有相同的保单限额，且没有免赔额。你需要决定各种不同总自留额及保单限额对应下的再保险费率，并且估计再保险赔付额的分布，获得分布的均值、标准差和 90% 以及 99% 分位数。

第三，研究是否需要对三个不同工作领域的董事和高级职员分开考虑，也就是检验三类董事和高级职员风险水平是否一致。你需要构建一个合适

① 数据来自 Klugman, S. T., H. H. Panjer and G.. E. Willmot "Loss Models: From Data to Decisions", John Wiley & Sons 1998，第 18～21 页。

的检验来判断三个不同工作领域的董事和高级职员的索赔额和索赔次数分布是否一致。是建立一个简单模型，还是建立三个不同模型？哪种更为合理？可以利用信度理论对此进行比较分析。更进一步，你预计客户可能要求经验定价。这将是一种追溯定价，即如果某年发生的实际再保险赔付很低，则下一年的再保险费将降低；反之，如果实际再保险赔付很高，则再保费将升高。

第四，你要说服你的客户，购买再保险是一个明智的选择。你可以通过比较五年内客户的破产概率来说明。如果购买了再保险，破产概率会降低。或者，你可以说明购买再保险将会使你的客户维持该项责任险的初始盈余降低。

表 14－4a　　保险公司董事和高级职员

年份	免赔额	最大支付额	赔付额	年份	免赔额	最大支付额	赔付额
1990	0	1 000 000	2 890	1991	15 000 000	10 000 000	10 000 000
1990	0	5 000 000	5 851	1992	0	1 000 000	1 836
1990	250 000	10 000 000	15 347	1992	0	1 000 000	10 705
1990	0	1 000 000	15 635	1992	0	5 000 000	10 973
1990	0	3 000 000	20 553	1992	0	5 000 000	13 408
1990	0	10 000 000	34 584	1992	0	10 000 000	16 339
1990	0	10 000 000	79 661	1992	350 000	5 000 000	95 736
1990	0	400 000	132 601	1992	0	1 000 000	212 313
1990	1 500 000	5 000 000	1 410 989	1992	0	5 000 000	439 543
1990	0	10 000 000	2 784 401	1992	70 000 000	15 000 000	1 098 710
1990	0	10 000 000	4 894 360	1992	0	3 000 000	1 211 180
1990	10 000 000	10 000 000	9 316 751	1993	0	500 000	10 510
1991	0	1 000 000	1 891	1993	0	3 000 000	14 029
1991	0	3 000 000	30 893	1993	0	10 000 000	15 296
1991	0	1 000 000	31 392	1993	50 000	1 000 000	27 516
1991	500 000	10 000 000	49 488	1993	0	10 000 000	53 467
1991	175 000	1 000 000	67 425	1993	300 000	5 000 000	87 463
1991	0	1 000 000	150 310	1993	100 000	5 000 000	220 995
1991	45 000 000	33 000 000	1 335 735	1993	150 000	5 000 000	274 086
1991	0	10 000 000	3 308 199	1993	0	5 000 000	1 862 304
1991	12 750 000	10 000 000	10 000 000	1993	0	5 000 000	5 000 000

表 14－4b　　银行董事和高级职员

年份	免赔额	最大支付额	赔付额	年份	免赔额	最大支付额	赔付额
1990	5 000	5 000 000	10 548	1991	0	10 000 000	4 435 099
1990	0	1 000 000	12 959	1991	0	5 000 000	5 000 000
1990	0	1 000 000	13 456	1991	10 000 000	20 000 000	5 644 894
1990	0	1 000 000	16 148	1992	0	5 000 000	1 003
1990	0	5 000 000	20 684	1992	0	1 000 000	2 388
1990	5 000 000	2 000 000	23 691	1992	0	5 000 000	3 067
1990	75 000	3 000 000	27 196	1992	0	10 000 000	4 066
1990	0	1 000 000	28 283	1992	0	10 000 000	6 758
1990	5 000 000	2 000 000	169 616	1992	0	1 000 000	6 781
1990	12 000 000	5 000 000	268 534	1992	0	3 000 000	7 439
1990	50 000	1 000 000	1 000 000	1992	1 000	5 000 000	10 617
1990	500 000	10 000 000	1 033 715	1992	1 000	5 000 000	10 888
1990	0	3 000 000	1 363 432	1992	0	7 500 000	34 745
1990	4 500 000	5 000 000	2 205 674	1992	0	400 000	58 587
1990	1 500 000	30 000 000	3 148 409	1992	0	1 000 000	113 166
1990	16 000 000	10 000 000	8 652 788	1992	0	5 000 000	122 967
1990	500 000	10 000 000	8 719 031	1992	350 000	3 000 000	199 607
1990	55 000 000	10 000 000	9 508 586	1992	150 000	5 000 000	298 847
1991	0	10 000 000	1 362	1992	75 000	1 000 000	1 000 000
1991	0	10 000 000	1 883	1992	0	1 000 000	1 000 000
1991	50 000	3 000 000	3 394	1992	0	10 000 000	3 022 258
1991	100 000	1 000 000	4 246	1992	0	5 000 000	3 201 434
1991	0	500 000	6 992	1992	10 000 000	10 000 000	3 754 944
1991	0	10 000 000	10 262	1992	10 000 000	5 000 000	5 000 000
1991	0	1 000 000	16 452	1992	700 000	20 000 000	6 126 080
1991	0	3 000 000	20 427	1993	0	5 000 000	189
1991	0	3 000 000	27 494	1993	0	10 000 000	388
1991	0	3 000 000	30 698	1993	0	1 000 000	2 026
1991	0	1 000 000	45 743	1993	0	1 000 000	2 354
1991	100 000	5 000 000	52 023	1993	0	10 000 000	8 959
1991	100 000	1 000 000	54 481	1993	0	1 000 000	17 865
1991	10 000 000	10 000 000	164 732	1993	0	10 000 000	41 170
1991	0	1 000 000	535 593	1993	0	5 000 000	158 391
1991	1 000 000	6 000 000	1 491 732	1993	1 000 000	10 000 000	596 674
1991	30 000 000	10 000 000	2 271 437	1993	100 000	10 000 000	926 657
1991	200 000	5 000 000	2 732 422	1993	1 000 000	10 000 000	1 101 816
1991	0	10 000 000	3 130 873	1993	100 000	5 000 000	1 903 358
1991	0	10 000 000	3 622 812	1993	0	10 000 000	2 055 117
1991	500 000	10 000 000	4 288 766	1993	0	10 000 000	2 966 399

表 14 - 4c　　证券公司董事和高级职员

年份	免赔额	最大支付额	赔付额
1990	0	7 500 000	60 664
1990	5 000 000	5 000 000	116 134
1990	0	7 500 000	576 857
1991	0	1 000 000	31 698
1991	50 000	10 000 000	46 427
1991	0	7 500 000	119 206
1991	450 000	1 000 000	405 796
1991	0	5 000 000	1 519 846
1991	500 000	2 000 000	2 000 000
1992	0	1 000 000	505
1992	250 000	5 000 000	17 833
1992	0	3 000 000	20 546
1992	22 000 000	10 000 000	10 000 000
1993	20 000	5 000 000	4 699
1993	150 000	10 000 000	6 055
1993	0	1 000 000	10 950
1993	25 000 000	10 000 000	244 026
1993	100 000	1 000 000	255 892
1993	0	5 000 000	384 222

表 14 - 5　　风险暴露数和损失次数

年份	保险公司		银行		证券公司	
	风险暴露数	损失次数	风险暴露数	损失次数	风险暴露数	损失次数
1990	853	20	1 446	27	639	5
1991	1 105	14	1 780	35	725	8
1992	1 148	16	1 717	36	685	4
1993	1 270	21	2 065	24	864	11

14.3.2　索赔强度建模

我们先对索赔强度进行建模分析。假设三类公司的董事和高级职员的损失强度分布都相同，用 X 表示一次损失事件的损失额。由于我们的数据有很多不同的免赔额和赔偿限额，而且极大似然估计具有很好的统计性质，因此我们选择极大似然估计来估计参数。如果保单的免赔额是 d，最大赔付额是 L，则获得最大赔付额的保单损失至少大于 $d+L$，因此我们记 $u=d$

+L 为最大损失。每次索赔事件，原保险公司的赔付额 Y^p 为：

$$Y^p=\begin{cases}未定义, & X<d\\ X-d, & d<X<u\\ u, & X>u\end{cases}$$

其分布密度函数为：

$$f_{Y^p}(x)=\begin{cases}\dfrac{f_X(x+d)}{1-F_X(d)}, & x<u\\ \dfrac{1-F_X(u)}{1-F_X(d)}, & x=u\end{cases} \tag{14.3.1}$$

其中，$f_X(x)$，$F_X(x)$ 分别是保单原损失 X 的分布密度函数和分布函数。我们初步选择了 12 个分布，根据式（14.3.1）写出极大似然函数，使用单纯形法，我们这里得到了 12 个分布的参数估计，见表 14－6。可以看出 12 个模型中对数正态分布得到的极大似然值最大（负的值最小）。参数估计结果为 $\hat{\mu}=10.5442$，$\hat{\sigma}=2.31307$。

表 14－6　　12 个分布的参数估计

分布名称	参数个数	参数估计	负极大似然值
Inverse exponential	1	$\theta=84.28567$	2 103.262
Lognormal	2	$\mu=10.544183$　$\sigma=2.313068$	1 781.173
韦伯	2	$\tau=0.1120341$　$\theta=8.5822086$	1 922.383
Loglogistic	2	$\gamma=0.2468978$　$\theta=14.0573170$	1 921.999
Paralogistic	2	$\alpha=0.088222$　$\theta=12.478940$	2 240.939
Inverse paralogistic	2	$\tau=0.0899498$　$\theta=18.2063624$	2 237.876
帕累托	2	$\alpha=0.1951758$　$\theta=30.6099376$	1 877.975
Inverse 帕累托	2	$\tau=38.22458$　$\theta=81.02645$	1 860.935
Inverse Gaussian	2	$\mu=1\,000\,005.82$　$\lambda=1\,016.012$	1 834.673
Inverse 韦伯	2	$\theta=20.5138478$　$\tau=0.2351234$	1 903.707
Inverse Burr	3	$\tau=0.0516133$　$\theta=12.5121518$　$\gamma=0.2104612$	2 229.544
Generalized 帕累托	3	$x_i=5.276204$　$\mu=38.441207$　$\beta=38.081655$	1 898.977

虽然从似然值上对数正态分布看起来比其他模型要优，但是我们还是需要做进一步的检验分析对数正态分布是否可以足够好地拟合样本数据，这样才可以放心地使用这个分布模型。否则的话，我们将不得不考虑其他一些复杂的模型，比如，两点混合分布。由于存在不同的免赔额，使得卡

方拟合优度检验无法使用。同样也没有明显的经验分布函数来做 Kolmogorov – Smirnov 检验。但是我们可以构造近似的经验分布函数，即 Kaplan – Meier 有限乘积估计量。

具体步骤如下。首先，对于每一个赔付 z_j，$j=1,\cdots,n$，免赔额 d_j，限额 u_j，损失 $x_j=d+z_j$。注意，在我们通常的定义下，最大赔付额是 u_j-d_j；在一般情况中，d_j 可能为 0，u_j 可能为无穷大。现在取出 $2n$ 个数 $\{d_1,\cdots,d_n,x_1,\cdots,x_n\}$，并将它们升序排列。如果两个或更多个 x_j 取值相同，$x_j=u_j$ 的数将放在 $x_j<u_j$ 的数之后。将这些排序后的数字 $t_j=d_k$ 标记为 $t_1,\cdots,t_{2n}$。接下来创建一系列数值 $\delta_1,\cdots,\delta_{2n}$，如果 t_j 等于某个 d_k，则 $\delta_j=0$；如果 t_j 等于某个 x_k 且 $x_k=u_k$，则 $\delta_j=1$；如果 t_j 等于某个 x_k 且 $x_k<u_k$，则 $\delta_j=2$。用公式表示为：

$$\delta_j=\begin{cases}0, & t_j=d_k\\1, & t_j=x_k,\ x_k=u_k\\2, & t_j=x_k,\ x_k<u_k\end{cases}$$

【例 14 – 1】 对表 14 – 4c 中的数据构造经验分布函数的 Kaplan – Meier 估计量。

解： 表 14 – 4 中有 19 个观察值。表 14 – 7 将其中的免赔、限额和赔付额转换成了 Kaplan – Meier 所要求的 d、u 和 x。表 14 – 8 将 $\{d_1,\cdots,d_n,x_1,\cdots,x_n\}$ 38 个元素从大到小的排列，按照上面的步骤将得到 38 个排序后的数值及它们原始数值。

表 14 – 7 证券公司董事和高级职员责任赔付

j	d_j	u_j	x_j
1	0	7 500 000	60 664
2	5 000 000	10 000 000	5 116 134
3	0	7 500 000	576 857
4	0	1 000 000	31 698
5	50 000	10 050 000	96 427
6	0	7 500 000	119 206
7	450 000	1 450 000	855 796
8	0	5 000 000	1 159 846
9	500 000	2 500 000	2 500 000
10	0	1 000 000	505
11	250 000	5 250 000	267 833
12	0	3 000 000	20 546
13	22 000 000	32 000 000	32 000 000

续表

j	d_j	u_j	x_j
14	20 000	5 020 000	24 699
15	150 000	10 150 000	156 055
16	0	1 000 000	10 950
17	25 000 000	35 000 000	25 244 023
18	100 000	1 100 000	355 892
19	0	5 000 000	384 222

表 14－8　　　　Kaplan－Meier 估计的排序数值

j	t_j	δ_j	j	t_j	δ_j
1	0	0	20	119 206	2
2	0	0	21	150 000	0
3	0	0	22	156 055	2
4	0	0	23	250 000	0
5	0	0	24	267 833	2
6	0	0	25	355 892	2
7	0	0	26	384 222	2
8	0	0	27	450 000	0
9	0	0	28	500 000	0
10	505	2	29	576 857	2
11	10 950	2	30	855 796	2
12	20 000	0	31	1 519 846	2
13	20 546	2	32	2 500 000	1
14	24 699	2	33	5 000 000	0
15	31 698	2	34	5 116 134	2
16	50 000	0	35	22 000 000	0
17	60 664	2	36	25 000 000	0
18	96 427	2	37	25 244 023	2
19	100 000	0	38	32 000 000	1

下面根据这些信息，构造 Kaplan－Meier 经验分布函数。从 $F_n(0)=0$ 开始。设置计数器 $r=s=0$，变量 $y_0=0$。对 $j=1,\cdots,2n$，进行如下操作：

- 如果 $\delta_j=0$，令 $r=r+1$；

- 如果 $\delta_j = 1$，令 $r = r - 1$；
- 如果 $\delta_j = 2$，令 $c = r - d$，$s = s + 1$，令 $y_s = t_j$，$F_n(y_s) = 1 - [1 - F_n(y_{s-1})] * c/r$，其中 d 为 t_j 相同且 $\delta_j = 2$ 的连续 js 的个数。然后跳到下一个$d - 1$ 项。

当完成上面的程序，用直线将 $F_n(y_s)$顺次连接起来便得到了经验分布。表 14 –9 是例 14 –2 的经验分布函数，图 14 –5 是经验分布函数图。

表 14 –9　　　　例 14 –2 中的 Kaplan – Meier 估计

j	r	s	y_s	$F_n(y_s)$	j	r	s	y_s	$F_n(y_s)$
1	1				20	4	8	119 026	0.7407
2	2				21	5			
3	3				22	4	9	156 055	0.7926
4	4				23	5			
5	5				24	4	10	267 833	0.8341
6	6				25	3	11	355 892	0.8756
7	7				26	2	12	384 222	0.9170
8	8				27	3			
9	9				28	4			
10	8	1	505	0.1111	29	3	13	576 857	0.9378
11	7	2	10 950	0.2222	30	2	14	855 796	0.9585
12	8				31	1	15	1 519 846	0.9793
13	7	3	20 546	0.3194	32	0			
14	6	4	24 699	0.4167	33	1			
15	5	5	31 698	0.5139	34	0	16	5 116 134	
16	6				35	1			
17	5	6	60 664	0.5949	36	2			
18	4	7	96 427	0.6759	37	1	17	25 244 023	
19	5				38	0			

注意，由于 r 在最后一个数值前达到了 0（最后一个数值总是 0），因此在 y_s 取值为 1 519 846 时，经验分布函数达到最大值 0.9793。只有当在组后一个观测是一个损失（$\delta = 2$）且 r 在此之前从未达到 0。

类似地，将三种类型的董事和高级职员的索赔额数据全部合在一起，根据上面的步骤求得经验分布函数。图 14 –5 中画出了的经验分布函数和拟合的对数正态分布。尽管拟合得不特别好，但已经相当不错了，特别是对于大损失的拟合（高于 exp（12.5）＝268 000）。

损失强度建模的最后任务是检查是否应当对三个样本分别建模，可使用似然比检验完成此项任务。零假设为三种类型的董事和高级职员有相同的损失强度分布，表 14 –10 给出了三个样本对应的极大似然估计值。似

然比检验统计量取值为 2 × (1 781.17 − 1 779.47) = 3.4。自由度为 4，p 值为 0.4932。因此，无法拒绝原假设。

表 14－10　三种类型的极大似然函数值

样　本	负极大似然函数值
保险	533.64
银行	1 015.37
证券	230.46
总和	1 779.47

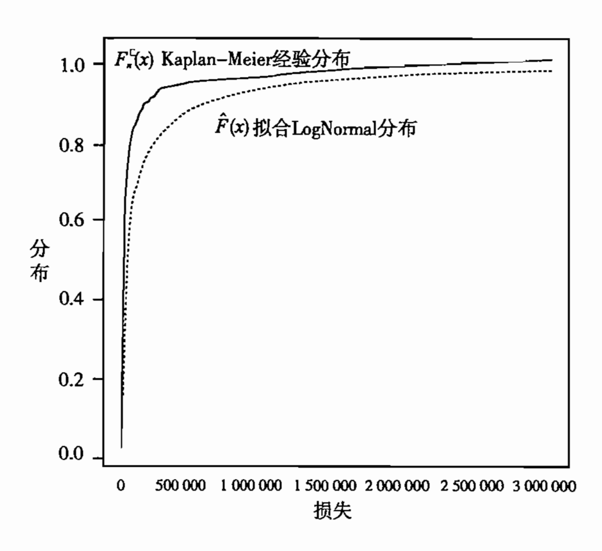


图14－5　证券公司董事和高级职员损失经验分布函数图

14.3.3　损失频率建模

对于损失频率的建模，有以下两种方法：

第一种方法：将这个混合的数据集看成是一个来自于无限可分总体的样本，其中每一个样本观测值的分布参数都不同，因为它们这些参数必须反映风险单位数，也就是说我们应该对每个样本的参数使用风险单位数进行调整。例如，如果单张保单损失次数服从参数为 λ 的泊松分布，则第 k 年保单年的第 i 类型董事和高级职员总损失次数的分布也服从泊松分布，只是泊松参数等于 λ 乘以风险单位数 $e_{k,i}$，即该类保单的签单数。因此，极大似然函数可以写为：

$$L = \prod_{k=1}^{4}\prod_{i=1}^{3}\frac{e^{-\lambda e_{k,i}}(\lambda e_{k,i})^{n_{k,i}}}{n_{k,i}!}$$

参数的极大似然估计值 $\hat{\lambda} = \sum_{k=1}^{4}\sum_{i=1}^{3} n_{k,i} / \sum_{k=1}^{4}\sum_{i=1}^{3} e_{k,i} = 0.0154578$，相应的负的对数极大似然函数值为 38.21111。

如果单个保单的损失次数服从负二项分布，参数为 β 和 r，且保单之间是相互独立的，则第 k 个保单年的第 i 类型年的损失次数也服从负二项分布，参数为 β 和 $re_{k,i}$。因此，极大似然函数为：

$$L = \prod_{k=1}^{4}\prod_{i=1}^{3}\frac{(re_{k,i}+n_{k,i}-1)!}{n_{k,i}!(re_{k,i}-1)!}\left(\frac{\beta}{1+\beta}\right)^{n_{k,i}}\left(\frac{1}{1+\beta}\right)^{re_{k,i}}$$

负二项分布参数的估计结果是 $\beta = 0.82842465$ 和 $r = 0.01857178$，相应的负的对数极大似然函数值为 36.91082。

其他分布的参数调整方法，我们在第四章已详细讨论。表 14－11 给出

了估计结果和相应的对数极大似然函数值。从表 14 – 11 可以看出，泊松分布是最优的分布。

第二种方法：假设保单每年最多发生一次损失，由于损失的频率是比较低的，因此这种假设是具有一定的合理性的。根据上面第一种方法得到的泊松分布的参数的估计值 0.0154587，损失次数大于 1 的概率为 $1-\exp(-0.0154587)-0.0154587\times\exp(-0.0154587)=0.0001182614$，大约为万分之一，非常小，可以忽略。因此这里只使用单参数的模型，泊松分布和几何分布。这里以泊松分布为例说明参数估计，几何分布的参数估计，请读者作为练习。

表 14 – 11　损失频率的参数估计值

分布	参　　数	负对数似然值
泊松	$\lambda=0.0154578$	38.21111
负二项	$\beta=0.82842465$，$r=0.01857178$	36.91082

由于损失次数 N 只取 0 和 1，且 $P(N=0)=\exp(-\lambda)$，$P(N=1)=\lambda\exp(-\lambda)$。似然函数为：

$$L=(\exp(-\lambda))^{n_0}\times(\lambda\exp(-\lambda))^{n_1}=\lambda^{n_1}\times\exp(-n\lambda)$$

其中 n_0 是没有发生损失的风险单位数，n_1 是发生一次损失的风险单位数，n 为样本量。根据样本数据，$n_0=221$，$n_1=14\ 076$，$n=14\ 297$。

参数 λ 的极大似然估计结果等于总的索赔次数除以总的风险单位数，即 $221/14\ 297=0.01545779$，相应的对数似然值为 $-1\ 142.491$。

两个分布参数估计结果和相应的对数极大似然函数值见表 14 – 12。可以看出，不管是采用哪一种方法，参数估计结果都显示泊松分布是较优的，且极大似然估计值均等于样本均值，因此，没有必要假定保单每年最多发生一次损失。

表 14 – 12　损失次数估计结果

分布	参　　数	负对数似然值
泊　松	$\lambda=0.0154578$	1 142.491
负二项	$\beta=0.01545779$	1 144.190

最后，既然我们已经选择了使用泊松分布来拟合损失次数，那么是对三组数据使用共同的模型，还是分开建立三个模型呢？构造似然比检验如下：

H_0：建立一个共同的模型更优

H_1：分开建立三个模型更优

表 14 – 13 给出了三个不同分布对应的极大似然函数值。似然比检验统计量的值 $=2(l_{Separate}-l_{All})=2(-33.5746-(-38.2111))=9.2732$，似然比检验统计量服从自由度等于 2 的 χ^2 分布，p 值等于 0.0097。检验的结果是拒绝原假设，也就是说这三组的频数分布之间还是有显著地差异的。

表 14－13　　三个不同分布对应的极大似然函数值

样　本		λ	负对数似然值
所有样本		0.0154578	38.2111
分组样本	保险	0.01622486	11.2885
	银行	0.0174087	13.6449
	证券	0.0096121	8.6412
	合计		33.5746

选择对三组数据建立一个共同的泊松分布模型，尽管三组数据中间频数分布的参数有差异，但是为了分析方便，仍然选择建立共同的模型。参数估计值 $\lambda=0.01545779$，相应的对数似然函数值为 $-1\ 142.491$。

14.3.4　总理赔额建模分析

通过前面的建模，我们得到精算责任险的个别损失额服从参数为 $\mu=10.5442$，$\sigma=2.31307$ 的对数正态分布。如果不考虑类型，损失次数服从参数为 $\lambda=0.0154578$ 的泊松分布。但是，实际上不同董事和高级职员责任险的损失次数分布的参数不一样。

下面，我们将在此基础上，求取不同再保险合约对应的再保险赔付额的分布。

1. 单个保单的再保险赔付额分布。考虑一个限额损失再保险，再保险人只赔偿超过自留额 d 的那部分原保单损失，最大赔付额为 $u-d$。其中 u 是原保单的赔偿限额。这时有两种可行方法来求再保险赔付额的分布。

第一种方法，先求出发生再保险赔付情况下赔付额 Y_P 的分布。这是一个混合分布，连续部分的概率密度函数为：

$$f_Y(x)=\frac{f_X(x+d)}{1-F_X(d)},\ 0\leqslant x<u-d$$

离散部分的概率为：

$$P(Y=u-d)=\frac{1-F_X(u)}{1-F_X(d)}$$

将这个分布离散化后，再用递推法或者快速傅里叶变换法（FFT）求再保险赔付额的分布。当然，这个时候再保险赔付次数的分布也要做出调整，它次数服从参数为 $\lambda(1-F_X(d))$ 的泊松分布。

第二种方法，可以先求出每次损失的再保险赔付额 Y_L 的分布。Y_L 的分布也是一个混合分布，连续部分的概率密度函数为：

$$f_Y(x)=f_X(x+d),\ 0\leqslant x<u-d$$

离散部分的概率为：

$$P(Y=0)=F_X(d)$$

$$P(Y=u-d)=1-F_X(u)$$

这种情况下，泊松分布的参数不变。

下面考虑对于不同的 d 和 u 组合后每张保单的损失分布。这里我们使用混合数据的频率分布。把 Y_L 的取值范围按步长为 10 000 分成各个区间段，然后使用舍入法得到 Y_L 的离散化分布，最后再使用递推法或者 FFT 法求总理赔的分布。或者将 Y_P 的分布做离散化，再求总理赔额的分布，经过计算得到两种方法求得的结果是一样的。

在所有的情况下，总理赔额分布的 90% 或 99% 分位点都为 0，说明多数情况下限额损失再保险是没有赔付的。这并不奇怪，因为不难求得出现零损失的概率是 $\exp(-0.0154578)=0.985$。当有免赔额后，出现 0 赔付的概率将会更高。表 4－14 给出了不同免赔额和赔偿限额的情况下，单个保单理赔额的均值、标准差、离散系数的统计情况。

表 14－14　　　　单个保单的限额损失再保险

免赔额（10^6）	赔偿限额（10^6）	均值	标准差	离散系数
0.5	1	778.5175	18 873.99	24.24350
0.5	5	2 910.492	94 619.16	32.50968
0.5	10	3 809.324	144 781.9	144 781.9
0.5	25	4 824.71	4 824.71	4 824.71
0.5	50	5 414.68	306 413.7	56.58945
1.0	5	2 131.974	80 382	37.70308
1.0	10	3 030.806	132 552	43.7349
1.0	25	4 046.269	219 521.9	54.25292
1.0	50	4 636.164	298 144.1	64.30837
5.0	10	898.8316	62 564.23	69.60617
5.0	25	1 914.294	162 499.5	84.8874
5.0	50	2 504.199	249 770.6	99.74071
10.0	25	1 015.463	111 064.6	109.3734
10.0	50	1 605.381	205 950.7	128.2877

毫无疑问，当免赔额或者赔偿限额增加时，风险（用离散系数来测度）将增加，而且只出售一张保单的风险是极大的。当然，只承保一个保单的组合是比较极端的情况，下面考虑多个保单组合的情况。

2. 包含 100 个保单的再保险保单组。考虑含有 100 份的再保险保单组合。假设所有保单的免赔额和限额都是相同的，那么求总理赔额的分布只

需要改变频率的分布。100 个相互独立的泊松随机变量的和仍然服从泊松分布，参数为原参数的 100 倍。重复单个保单情况下的计算过程，我们可以得到该保单组合在不同免赔额和赔偿限额情况下的各种数据，结果见表 14－15。

显然，100 份相互独立的保单组，其均值是单个保单均值的 100 倍，标准差为单份保单的 10 倍。这意味着，离散系数将会是单份保单的 1/10。在所有情况下，99% 的分位数均大于 0。这表明风险增加了，但在实际中，这说明再保险赔付的可能性增大了。

表 14－15　　100 个保单的限额损失再保险

免赔额 (10^6)	赔偿限额 (10^6)	均值	标准差	离散系数	90% 分位数	99% 分位数
0.5	5	291 049	946 192	3.251	710 000	4 500 000
0.5	10	380 932	1 447 819	3.801	710 000	9 500 000
0.5	25	482 471	48 247	0.100	710 000	11 670 000
1	5	213 197	803 820	3.770	190 000	4e +06
1	10	303 081	1 325 520	4.373	190 000	9e +06
1	25	404 627	2 195 219	5.425	190 000	11 080 000
5	10	89 883	625 642	6.961	0	5e +06
5	25	191 429	1 624 995	8.489	0	6 890 000
10	25	101 546	1 110 646	10.937	0	1 850 000

3. 包含 100 份保单的总限额损失再保险。现在考虑对总损失的再保险。假设各份保单没有个体免赔额，但存在限额 u，也即 $d=0$，$u>0$。对这 100 个保单组合，再保险人只赔偿超过总自留额 a 的那部分总损失。与前面一样，先对损失额分布进行修正和离散化，泊松参数乘以 100，然后使用递推公式或者 FFT 方法获得总损失分布的离散分布。令其累积分布函数为 $F_S(s)$，概率函数为 $f_S(s_i)$，$i=1,\cdots,n$。如果总自留额是 a，那么再保险赔付 S_r 的分布函数为：

$$F_{S_r}(s) = F_S(s+a), \quad s \geqslant 0$$

$$f_{S_r}(0) = F_S(a) = \sum_{s_i \leqslant a} f_S(s_i)$$

$$f_{S_r}(r_i) = f_S(r_i + a), \qquad r_i = s_i - a, \quad i = 1,\cdots,n$$

取步长为 10 000，给定不同的总自留额和赔偿限额时，得到的统计结果见表 14－16。这个结果同限额损失再保险比较相似。大多数情况下，当个体的赔偿限额或者总自留额增加时，用离散系数测度的风险也增加。唯一的例外就是个体赔偿限额和总体免赔额都是 5 000 000 时的情况。这种设

置风险很大，因为这是唯一在再保险生效前发生两个损失的情况。

表 14－16　　包含 100 份保单的总限额损失再保险赔付分布

免赔额	赔偿限额	均值	标准差	离散系数	90%分位数	99%分位数
0.5	5	323 813.9	1 056 515	3.262722	877 692.8	4 726 496
0.5	10	413 697.1	1 553 935	3.756213	877 692.8	9 518 528
0.5	25	515 243.4	2 389 120	4.636877	877 692.8	11 910 237
1	5	242 682.8	914 446.4	3.768073	377 692.8	4 226 496
1	10	332 565.9	1 429 971	4.299813	377 692.8	9 018 528
1	25	434 112.2	2 288 330	5.271287	377 692.8	11 410 237
2.5	5	114 977.0	569 788.2	4.95567	0	2 726 496
2.5	10	204 860.2	1 124 106	5.487187	0	7 518 528
2.5	25	306 406.5	2 037 305	6.649027	0	9 910 237
5	5	13 476.9	182 324.5	13.52867	0	226 497
5	10	103 360.1	723 037.1	6.995324	0	5 018 528
5	25	204 906.3	1 703 540	8.31375	0	7 410 237

现假设这 100 个保单分为三类，损失次数泊松参数不同（但有相同的损失额分布）。第一组是 30 张保险公司的董事和高级职员责任险保单，单张保单的损失次数的泊松参数 $\lambda=0.0162249$，因此这 30 张的保单组合损失次数构成的组服从泊松分布，均值为 $30(0.0162249)=0.486747$。第二组是 50 张银行的董事和高级职员责任险保单，保单组合的泊松参数为 $50(0.0174087)=0.870435$。第三组是 20 张证券公司的董事和高级职员责任险保单。保单组合的泊松参数为 $20(0.0096121)=0.192242$。有三种方法可以构造这三组之和的分布。

（1） 因为独立泊松分布的和也是泊松分布，泊松参数为 1.549424。一般损失量分布还是对数正态分布。简化成一个复合分布，可以用任意方法估计。

（2） 分别得到三个聚合分布。如果利用递推或是 FFT 方法得到三个离散分布，可以用卷积得到和的分布。

（3） 如果用 FFT 或是 Heckman－Meyers 方法得到三个变换然后再相乘，给出乘积后的逆变换形式。

每种方法都各有优缺点。第一种方法要求其和具有已知形式的频率分布。如果索赔额分布不同，就无法将其合并形成一个模型。它的主要优点是，当模型适用时只需进行一次聚合计算。

第二种方法的优点在于它对于频率分布和损失程度分布没有限制。缺点是计算机存储量的快速膨胀，例如，第一个分布需要 3 000 个点，第二个

分布需要 5 000 个点，第三个分布需要 2 000 个点（三个分布的离散化区间相同），则联合分布将有 10 000 个点。

第三种方法对于三个模型也没有要求。它的缺点和第二种方法相同，但是存储的膨胀是在合并之前发生的，即三个分布都要计算 10 000 个点。似乎没有办法避免这一点。

14.3.5 信度调整与经验定价

我们已经知道，三种不同类型的董事和高级职员责任险的个别损失额分布相同，损失次数都服从泊松分布，但是参数不相同。但是由于经验数据太少，估计值的可信度不高。下面用信度理论对估计值进行调整。在此基础上，我们将制订一个经验费率计划。

信度理论可以用来改善索赔强度、索赔频率、总索赔额期望值的估计。在这里，我们将信度理论运用于改善期望索赔频率的估计值。这是因为，首先，由于这三类责任险保单的损失强度的分布是相同的，因此观察到的索赔额的随机波动不会影响到我们的最终结果；其次，由于我们的目标是分析不同的再保险安排的效果，将索赔频率和索赔强度分开考虑是非常有用的，因此我们只需要考虑索赔频率的信度估计。总共有四种信度调整方法来改善索赔频率的估计。

1. *有限波动信度*。设 N_i，$i=1$，2，3 表示三类保单一个风险单位在一年内发生的损失次数，它们都服从泊松分布，参数分别为 λ_i，$i=1$，2，3。N_{i1}，N_{i2}，…，N_{in}为第 i 类保单在过去相互独立的 n 个观察期内发生的索赔次数。首先假设 N_i，$i=1$，2，3 是独立同分布的，则 $\lambda_i=\lambda$，$i=1$，2，3。不妨用 N 表示一个风险单位在一年内发生的损失次数，服从泊松分布。参数为 λ。表 14－5 可看做 N 在观察期的观测值。下面来确定这些观测值是否完全可信。

首先，由第十二章知，对于泊松分布来说，保证经验数据完全可信所需的最小总索赔数 $\lambda_0=(y_p/r)^2$，当 $r=0.05$，$p=0.9$ 时，经验数据完全可信所需的最小总索赔数为 1 082.41，所需的最小风险单位 1 082.41/λ。也就是说，在所有的观察期内至少有 1 083 个索赔事件发生，或者至少有 1 082.41/λ 个风险单位的观测值，才能使索赔频率估计值完全可信。因为我们不知道 λ 的具体值，我们通过经验总体的均值来估计 λ，即 $\hat{\lambda}=221/14\ 297=0.015458$。因此完全信度条件所需要的最小风险单位数为 1 082.41/$\hat{\lambda}$ = 70 023.60。显然，我们观测到的风险单位数 14 297 <70 023.6，因此经验数据没有达到完全可信条件。

其次，由于经验数据是部分可信的，因此需要进行信度调整，信度因子为 $Z=(\lambda n/1\ 082.41)^{1/2}$。如果信度是基于损失次数，则用观察到的损失

次数来代替 λn；如果信度是基于风险单位数的，那么用 $\hat{\lambda}=221/14\ 297=0.015458$ 来代替，用总风险单位数来代替 n。估计出信度因子 Z 后，用分组数据的经验估计值与全体数据的经验均值的信度加权值来作为三类索赔频率的信度估计，即 $Z\bar{X}_i+(1-Z)\mu$，其中 $\bar{X}_i=c_i/m_i$，$\mu=0.015458$。c_i 表示第 i 组的损失次数，m_i 表示第 i 组的风险暴露数，具体的计算结果如表 14－17 所示。

表 14－17　　有限波动信度估计

组	风险暴露数	索赔数	索赔数/风险暴露数	基于损失数		基于风险暴露数	
				信度因子	λ_i	信度因子	λ_i
1	4 376	71	0.01622	0.2561	0.01565	0.2500	0.01565
2	7 008	122	0.01741	0.3357	0.01611	0.3164	0.01608
3	2 913	28	0.00961	0.1608	0.01452	0.2040	0.01427

最后，注意到在上述两种估计存在冲销误差，即根据参数的信度估计值算出来的预计损失数不等于实际发生的损失数。例如，假设参数的信度估计值基于损失次数，则在观察内应发生的损失次数应为 $\sum_{i=1}^{3} m_i\hat{\lambda}_i=223.71$，不等于实际发生的损失次数 221。因此，我们需要对估计值进行冲销调整，调整因子为 $221/223.71=0.98789$。将三类保单的信度估计值 0.01565，0.01611，0.01452 乘以调整因子，调整为 0.01546，0.01592，0.01434。同样的，基于风险单位数的期望索赔数为 222.69，调整因子为 0.99240，λ 的估计值由 0.01565，0.01608，0.01427 调整为 0.01553，0.01596，0.01416。

2. 经验贝叶斯信度。对于经验贝叶斯估计，不需要假设损失分布。在这里，m_{ij} 表示第 i 组第 j 年的风险单位数。第 i 组第 j 年的每风险单位损失频率 $X_{ij}=c_{ij}/m_{ij}$，第 i 组的平均损失频率 $\bar{X}_i=c_i/m_i$。从表 14－5 我们得知 $n_i=4$，$r=3$，$c=221$，$m=14\ 297$，可以计算：

$$
\begin{aligned}
\hat{v} &= \frac{1}{9}\sum_{i=1}^{3}\sum_{j=1}^{4} m_{ij}(c_{ij}/m_{ij}-c_i/m_i)^2 \\
&= \frac{1}{9}\sum_{i=1}^{3}\sum_{j=1}^{4} c_{ij}^2/m_{ij}-2c_{ij}c_i/m_i+m_{ij}c_i^2/m_i^2 \\
&= \frac{1}{9}\left(\sum_{i=1}^{3}\sum_{j=1}^{4} c_{ij}^2/m_{ij}-\sum_{i=1}^{3} c_i^2/m_i\right) \\
&= \frac{1}{9}(3.733444-3.544962)=0.0209424
\end{aligned}
$$

$$
\hat{a}=\left(m-m^{-1}\sum_{i=1}^{3} m_i^2\right)^{-1}\left[\sum_{i=1}^{3} m_i(c_i/m_i-c/m)^2-\hat{v}(3-1)\right]
$$

$$= (8\ 928.95)^{-1}\left(\sum_{i=1}^{3} c_i^2/m_i - c^2/m - 0.0418848\right)$$

$$= (8\ 928.95)^{-1}(0.086906) = 0.00000973306$$

所以得到 $\hat{k} = \hat{v}/\hat{a} = 2\ 151.68$，信度因子为 $Z_i = m_i/(2\ 151.68 + m_i)$。根据式(12.5.25)，信度加权后整体均值为 $\hat{\mu} = 0.02972/2.0107 = 0.01478$。表 14－18 给出了各组保单的信度估计值的计算过程。其中最后一列 $\hat{\lambda}_i = Z_i\bar{X}_i + (1 - Z_i)\hat{\mu}$ 注意，由于 $\sum m_i\hat{\lambda}_i = 221$，因此不需要对结果进行调整。

表 14－18　　经验贝叶斯信度估计

组	m_i	Z_i	$\bar{X}_i$	$Z_i\bar{X}_i$	$\hat{\lambda}_i$
1	4 376	0.6704	0.01622	0.01087	0.01575
2	7 008	0.7651	0.01741	0.01332	0.01679
3	2 913	0.5752	0.00961	0.00553	0.01181
合计		2.0107		0.02972	

3. 半参数信度。在损失次数的建模时，我们已经知道泊松分布是适合于各类责任险损失次数的，但泊松参数的先验分布未知。

假设各组保单的对应泊松参数的先验分布是稳定的，而且组内的每张保单的泊松参数是相同的。因此，第 i 组的损失次数 c_i 服从参数为 $m_i\lambda_i$ 的泊松分布。那么信度公式中几个关键变量分别为：

$$\mu = E\left(\frac{c_i}{m_i}\right) = E\left[E\left(\frac{c_i}{m_i}\middle|\lambda_i\right)\right] = E(\lambda_i)$$

$$v_i = E\left[Var\left(\frac{c_i}{m_i}\middle|\lambda_i\right)\right] = E\left(\frac{\lambda_i}{m_i}\right) = \mu/m_i$$

$$a = Var\left[E\left(\frac{c_i}{m_i}\middle|\lambda_i\right)\right] = Var(\lambda_i) = \sigma^2$$

$$Z_i = \frac{a}{a + v_i}$$

令 $\bar{X} = \left(\sum_{i=1}^{3} c_i/\sum_{i=1}^{3} m_i\right)$，由于

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{3} c_i}{\sum_{i=1}^{3} m_i}\right) = E\left(E\left(\frac{\sum_{i=1}^{3} c_i}{\sum_{i=1}^{3} m_i}\middle|\lambda_1, \lambda_2, \lambda_3\right)\right)$$

$$= E\left(\frac{\sum_{i=1}^{3} m_i\lambda_i}{\sum_{i=1}^{3} m_i}\right) = E(\lambda_i) = \mu$$

因此，$\hat{\mu} = 221/14\ 297 = 0.015458$。定义

$$SS = \sum_{i=1}^{3} m_i\left(\frac{c_i}{m_i} - \bar{X}\right)^2 = \sum_{i=1}^{3}\frac{c_i^2}{m_i} - \frac{\left(\sum_{i=1}^{3} c_i\right)^2}{\sum_{i=1}^{3} m_i}$$

其期望值等于

$$E(SS) = E\left[E\left(\sum_{i=1}^{3}\frac{c_i^2}{m_i} - \frac{\left(\sum_{i=1}^{3} c_i\right)^2}{\sum_{i=1}^{3} m_i}\middle|\lambda_1, \lambda_2, \lambda_3\right)\right]$$

$$= E\left(\sum_{i=1}^{3}\frac{m_i\lambda_i + m_i^2\lambda_i^2}{m_i} - \frac{\sum_{i=1}^{3}(m_i\lambda_i + m_i^2\lambda_i^2) + 2\sum_{i<j} m_i m_j \lambda_i \lambda_j}{\sum_{i=1}^{3} m_i}\right)$$

$$= 3\mu + \left(\sum_{i=1}^{3} m_i\right)(\sigma^2 + \mu^2) - \mu - \frac{1}{\sum_{i=1}^{3} m_i}\left[\left(\sum_{i=1}^{3} m_i^2\right)(\sigma^2 + \mu^2) + 2\sum_{i<j} m_i m_j \mu^2\right] \qquad (14.3.1)$$

$$= 2\mu + \left(\sum_{i=1}^{3} m_i\right)(\sigma^2 + \mu^2) - \sigma^2\frac{\sum_{i=1}^{3} m_i^2}{\sum_{i=1}^{3} m_i} - \mu^2\sum_{i=1}^{3} m_i$$

$$= 2\mu + \sigma^2\left(\sum m_i - \frac{\sum m_i^2}{\sum m_i}\right)$$

由样本数据计算得 $SS = 0.128791$。根据式（14.3.1）可得 $0.128791 = 2(0.015458) + \sigma^2(8\ 928.95)$，于是 $\sigma^2 = 0.0000109616$。因此

$$Z_i = \frac{a}{a + v_i} = \frac{\sigma^2}{\sigma^2 + \mu/m_i} = \frac{0.0000109616}{0.0000109616 + 0.015458/m_i} = \frac{m_i}{m_i + 1\ 410.18}$$

表 14－19 给出了具体计算结果。

表 14－19　　半参数信度估计

组	m_i	Z_i	$\bar{X}_i$	$Z_i\bar{X}_i$	$\hat{\lambda}_i$
1	4 376	0.7563	0.01622	0.01227	0.01585
2	7 008	0.8325	0.01741	0.01449	0.01695
3	2 913	0.6738	0.00961	0.00648	0.01127
合计		2.2626		0.03324	

4. 参数估计。为了简单起见，我们假定泊松参数 λ 服从均值为 μ 的指数分布。根据最大似然估计，我们可以得到下面的等式：

$$3\mu = \sum_{i=1}^{3}\frac{c_i + 1}{m_i + \mu^{-1}}$$

可以得出估计 $\hat{\mu}=0.014434$，我们可以得出 λ_i 的后验估计：

$$\hat{\lambda}_i=\frac{c_i+1}{m_i+\mu^{-1}}$$

所以这三组的估计值分别为0.01620，0.01738，0.00972。

5. 经验定价。FW公司最后的要求是制订一个经验费率计划。许多购买保险的董事和高级职员认为他们不大可能发生损失，而且认为缴纳的保费过高。然而，必须等到一年后才能知道损失是否发生以及损失发生的总金额，从而证明他们的风险水平确实很低。采用经验费率可以解决这个问题，即根据董事和高级职员当年的损失来确定下一个年度的费率。如果损失没有发生，则下一年的费率降低，反之则提高。

由于样本数据中221个损失事件无法说明一个董事和高级职员是否在一年内发生两次或三次事件，因此半参数估计的方法不适用。唯一可以使用的方法是全参数方法。例如，对于保险公司这一组数据，假设泊松参数服从均值为 μ 的指数分布，索赔频率的先验均值的估计值等于

$$\mu=\frac{71+1}{4\ 376+\mu^{-1}}$$

于是得 $\mu=71/4\ 376=0.01622$。如果某个董事和高级职员在一年中有 c 次索赔，则他的下一年经验费率应该根据期望索赔数

$$\hat{\lambda}=\frac{c+1}{1+0.01622^{-1}}=\frac{c+1}{62.6338}$$

来计算。对于其他两组数据来说，经验费率公式稍微有些不同，主要区别在于根据经验来计算各组的 μ 不同。

14.3.6 偿付能力分析

最后，在对各种再保险安排计划比较分析后，我们将试图说服原保险公司，购买再保险是一个明智的选择。为此需要进行偿付能力分析，证明再保险后破产概率将会减小。我们假设100个董事和高级职员是随机选定的，所以总的损失次数服从参数为1.54578的泊松分布；在前面又已知所有的董事和高级职员责任险损失额服从参数为 $\mu=10.54$ 和 $\sigma=2.31$ 的对数正态分布；我们还假设所有保单都没有免赔额，保单的赔偿限额是1 000万元。出于对风险的考虑，原保险人在净保费的基础上加了20%的安全附加。现在我们想要提供一个自留额为100万元的限额损失再保险，也就是说再保险公司对每次损失超过100万元以上的那部分进行赔付。再保险保费是再保险赔付的期望值的125%。为了使原保险人相信会从这个安排中受益，只需要说明如果购买了再保险，那么5年内发生破产的概率 $\tilde{\psi}(u,5)$ 会变小。

$$原保险人收到的总保费 = 1.2 \times 1.154578 \times E(X \wedge 10\ 000\ 000) = 697\ 980\ (元)$$

将损失额的分布离散化，区间个数为 10 000，再使用递推法可以得到总赔付额的分布。假设没有利息收入，使用卷积的方法，那么可以得到初始资金为 100 万元时，没有再保险安排时，5 年内的破产概率是 0.268。

$$再保险人的再保费收入 = 1.25 \times 1.54578 \times [E(X \wedge 10\ 000\ 000) - E(X \wedge 1\ 000\ 000)] = 378\ 851\ (元)$$

原保险人自留的保费收入为 319 129 元。由于原保险公司的赔付额低于 100 万元，重新使用递推法获得总赔付额的分布，利用卷积计算，破产概率减少到 0.176。

可以从另外一个角度看再保险的好处：如果将原始资金降到 70 万元，那么这时的破产概率为 0.271，与没有再保险的破产概率差不多，这可能意味着再保险并不有效。为了达到减少 30 万元初始资金的要求，原保险人需放弃 378 851 元的保费收入。这看起来不是一个好主意。

附　录

附录一　中国人寿保险业经验生命表

中国人寿保险业经验生命表（2000—2003）（男）						
China Life Insurance Mortality Table（2000—2003）						
非养老金业务男表 CL1（2000—2003）						
年龄	死亡率	生存人数	死亡人数	生存人年数		平均余命
(x)	q_x	l_x	d_x	L_x	T_x	$\mathring{e}_x$
0	0.000722	1 000 000	722	999 639	76 712 498	76.71
1	0.000603	999 278	603	998 977	75 712 859	75.77
2	0.000499	998 675	498	998 426	74 713 882	74.81
3	0.000416	998 177	415	997 970	73 715 456	73.85
4	0.000358	997 762	357	997 584	72 717 486	72.88
5	0.000323	997 405	322	997 244	71 719 902	71.91
6	0.000309	997 083	308	996 929	70 722 658	70.93
7	0.000308	996 775	307	996 622	69 725 729	69.95
8	0.000311	996 468	310	996 313	68 729 107	68.97
9	0.000312	996 158	311	996 003	67 732 794	67.99
10	0.000312	995 847	311	995 692	66 736 791	67.02
11	0.000312	995 536	311	995 381	65 741 099	66.04
12	0.000313	995 225	312	995 069	64 745 718	65.06
13	0.00032	994 913	318	994 754	63 750 649	64.08
14	0.000336	994 595	334	994 428	62 755 895	63.1
15	0.000364	994 261	362	994 080	61 761 467	62.12
16	0.000404	993 899	402	993 698	60 767 387	61.14
17	0.000455	993 497	452	993 271	59 773 689	60.16
18	0.000513	993 045	509	992 791	58 780 418	59.19
19	0.000572	992 536	568	992 252	57 787 627	58.22
20	0.000621	991 968	616	991 660	56 795 375	57.26

续表

中国人寿保险业经验生命表（2000—2003）（男）						
China Life Insurance Mortality Table（2000—2003）						
非养老金业务男表 CL1（2000—2003）						
年龄	死亡率	生存人数	死亡人数	生存人年数		平均余命
(x)	q_x	l_x	d_x	L_x	T_x	$\mathring{e}_x$
21	0.000661	991 352	655	991 025	55 803 715	56.29
22	0.000692	990 697	686	990 354	54 812 690	55.33
23	0.000716	990 011	709	989 657	53 822 336	54.37
24	0.000738	989 302	730	988 937	52 832 679	53.4
25	0.000759	988 572	750	988 197	51 843 742	52.44
26	0.000779	987 822	770	987 437	50 855 545	51.48
27	0.000795	987 052	785	986 660	49 868 108	50.52
28	0.000815	986 267	804	985 865	48 881 448	49.56
29	0.000842	985 463	830	985 048	47 895 583	48.6
30	0.000881	984 633	867	984 200	46 910 535	47.64
31	0.000932	983 766	917	983 308	45 926 335	46.68
32	0.000994	982 849	977	982 361	44 943 027	45.73
33	0.001055	981 872	1 036	981 354	43 960 666	44.77
34	0.001121	980 836	1 100	980 286	42 979 312	43.82
35	0.001194	979 736	1 170	979 151	41 999 026	42.87
36	0.001275	978 566	1 248	977 942	41 019 875	41.92
37	0.001367	977 318	1 336	976 650	40 041 933	40.97
38	0.001472	975 982	1 437	975 264	39 065 283	40.03
39	0.001589	974 545	1 549	973 771	38 090 019	39.08
40	0.001715	972 996	1 669	972 162	37 116 248	38.15
41	0.001845	971 327	1 792	970 431	36 144 086	37.21
42	0.001978	969 535	1 918	968 576	35 173 655	36.28
43	0.002113	967 617	2 045	966 595	34 205 079	35.35
44	0.002255	965 572	2 177	964 484	33 238 484	34.42
45	0.002413	963 395	2 325	962 233	32 274 000	33.5
46	0.002595	961 070	2 494	959 823	31 311 767	32.58
47	0.002805	958 576	2 689	957 232	30 351 944	31.66
48	0.003042	955 887	2 908	954 433	29 394 712	30.75
49	0.003299	952 979	3 144	951 407	28 440 279	29.84

续表

中国人寿保险业经验生命表（2000—2003）（男）						
China Life Insurance Mortality Table（2000—2003）						
非养老金业务男表 CL1（2000—2003）						
年龄	死亡率	生存人数	死亡人数	生存人年数		平均余命
(x)	q_x	l_x	d_x	L_x	T_x	$\mathring{e}_x$
50	0.00357	949 835	3 391	948 140	27 488 872	28.94
51	0.003847	946 444	3 641	944 624	26 540 732	28.04
52	0.004132	942 803	3 896	940 855	25 596 108	27.15
53	0.004434	938 907	4 163	936 826	24 655 253	26.26
54	0.004778	934 744	4 466	932 511	23 718 427	25.37
55	0.005203	930 278	4 840	927 858	22 785 916	24.49
56	0.005744	925 438	5 316	922 780	21 858 058	23.62
57	0.006427	920 122	5 914	917 165	20 935 278	22.75
58	0.00726	914 208	6 637	910 890	20 018 113	21.9
59	0.008229	907 571	7 468	903 837	19 107 223	21.05
60	0.009313	900 103	8 383	895 912	18 203 386	20.22
61	0.01049	891 720	9 354	887 043	17 307 474	19.41
62	0.011747	882 366	10 365	877 184	16 420 431	18.61
63	0.013091	872 001	11 415	866 294	15 543 247	17.82
64	0.014542	860 586	12 515	854 329	14 676 953	17.05
65	0.016134	848 071	13 683	841 230	13 822 624	16.3
66	0.017905	834 388	14 940	826 918	12 981 394	15.56
67	0.019886	819 448	16 296	811 300	12 154 476	14.83
68	0.022103	803 152	17 752	794 276	11 343 176	14.12
69	0.024571	785 400	19 298	775 751	10 548 900	13.43
70	0.027309	766 102	20 921	755 642	9 773 149	12.76
71	0.03034	745 181	22 609	733 877	9 017 507	12.1
72	0.033684	722 572	24 339	710 403	8 283 630	11.46
73	0.037371	698 233	26 094	685 186	7 573 227	10.85
74	0.04143	672 139	27 847	658 216	6 888 041	10.25
75	0.045902	644 292	29 574	629 505	6 229 825	9.67
76	0.050829	614 718	31 246	599 095	5 600 320	9.11
77	0.056262	583 472	32 827	567 059	5 001 225	8.57
78	0.062257	550 645	34 282	533 504	4 434 166	8.05

续表

中国人寿保险业经验生命表（2000—2003）（男）						
China Life Insurance Mortality Table（2000—2003）						
非养老金业务男表 CL1（2000—2003）						
年龄	死亡率	生存人数	死亡人数	生存人年数		平均余命
(x)	q_x	l_x	d_x	L_x	T_x	$^\circ e_x$
79	0.068871	516 363	35 562	498 582	3 900 662	7.55
80	0.076187	480 801	36 631	462 486	3 402 080	7.08
81	0.084224	444 170	37 410	425 465	2 939 594	6.62
82	0.093071	406 760	37 858	387 831	2 514 129	6.18
83	0.1028	368 902	37 923	349 941	2 126 298	5.76
84	0.113489	330 979	37 562	312 198	1 776 357	5.37
85	0.125221	293 417	36 742	275 046	1 464 159	4.99
86	0.13808	256 675	35 442	238 954	1 189 113	4.63
87	0.152157	221 233	33 662	204 402	950 159	4.29
88	0.167543	187 571	31 426	171 858	745 757	3.98
89	0.184333	156 145	28 783	141 754	573 899	3.68
90	0.202621	127 362	25 806	114 459	432 145	3.39
91	0.2225	101 556	22 596	90 258	317 686	3.13
92	0.244059	78 960	19 271	69 325	227 428	2.88
93	0.267383	59 689	15 960	51 709	158 103	2.65
94	0.292544	43 729	12 793	37 333	106 394	2.43
95	0.319604	30 936	9 887	25 993	69 061	2.23
96	0.348606	21 049	7 338	17 380	43 068	2.05
97	0.379572	13 711	5 204	11 109	25 688	1.87
98	0.412495	8 507	3 509	6 753	14 579	1.71
99	0.447334	4 998	2 236	3 880	7 826	1.57
100	0.48401	2 762	1 337	2 094	3 946	1.43
101	0.522397	1 425	744	1 053	1 852	1.3
102	0.562317	681	383	490	799	1.17
103	0.603539	298	180	208	309	1.04
104	0.64577	118	76	80	101	0.86
105	1	42	42	21	21	0.5

中国人寿保险业经验生命表（2000—2003）（女）						
China Life Insurance Mortality Table（2000—2003）						
非养老金业务女表 CL2（2000—2003）						
年龄	死亡率	生存人数	死亡人数	生存人年数		平均余命
(x)	q_x	l_x	d_x	L_x	T_x	$\mathring{e}_x$
0	0.000661	1 000 000	661	999 670	80 892 053	80.89
1	0.000536	999 339	536	999 071	79 892 383	79.95
2	0.000424	998 803	423	998 592	78 893 312	78.99
3	0.000333	998 380	332	998 214	77 894 720	78.02
4	0.000267	998 048	266	997 915	76 896 506	77.05
5	0.000224	997 782	224	997 670	75 898 591	76.07
6	0.000201	997 558	201	997 458	74 900 921	75.08
7	0.000189	997 357	189	997 263	73 903 463	74.1
8	0.000181	997 168	180	997 078	72 906 200	73.11
9	0.000175	996 988	174	996 901	71 909 122	72.13
10	0.000169	996 814	168	996 730	70 912 221	71.14
11	0.000165	996 646	164	996 564	69 915 491	70.15
12	0.000165	996 482	164	996 400	68 918 927	69.16
13	0.000169	996 318	168	996 234	67 922 527	68.17
14	0.000179	996 150	178	996 061	66 926 293	67.18
15	0.000192	995 972	191	995 877	65 930 232	66.2
16	0.000208	995 781	207	995 678	64 934 355	65.21
17	0.000226	995 574	225	995 462	63 938 677	64.22
18	0.000245	995 349	244	995 227	62 943 215	63.24
19	0.000264	995 105	263	994 974	61 947 988	62.25
20	0.000283	994 842	282	994 701	60 953 014	61.27
21	0.0003	994 560	298	994 411	59 958 313	60.29
22	0.000315	994 262	313	994 106	58 963 902	59.3
23	0.000328	993 949	326	993 786	57 969 796	58.32
24	0.000338	993 623	336	993 455	56 976 010	57.34
25	0.000347	993 287	345	993 115	55 982 555	56.36
26	0.000355	992 942	352	992 766	54 989 440	55.38

续表

中国人寿保险业经验生命表（2000—2003）（女）						
China Life Insurance Mortality Table（2000—2003）						
非养老金业务女表 CL2（2000—2003）						
年龄	死亡率	生存人数	死亡人数	生存人年数		平均余命
(x)	q_x	l_x	d_x	L_x	T_x	$°e_x$
27	0.000362	992 590	359	992 411	53 996 674	54.4
28	0.000372	992 231	369	992 047	53 004 263	53.42
29	0.000386	991 862	383	991 671	52 012 216	52.44
30	0.000406	991 479	403	991 278	51 020 545	51.46
31	0.000432	991 076	428	990 862	50 029 267	50.48
32	0.000465	990 648	461	990 418	49 038 405	49.5
33	0.000496	990 187	491	989 942	48 047 987	48.52
34	0.000528	989 696	523	989 435	47 058 045	47.55
35	0.000563	989 173	557	988 895	46 068 610	46.57
36	0.000601	988 616	594	988 319	45 079 715	45.6
37	0.000646	988 022	638	987 703	44 091 396	44.63
38	0.000699	987 384	690	987 039	43 103 693	43.65
39	0.000761	986 694	751	986 319	42 116 654	42.68
40	0.000828	985 943	816	985 535	41 130 335	41.72
41	0.000897	985 127	884	984 685	40 144 800	40.75
42	0.000966	984 243	951	983 768	39 160 115	39.79
43	0.001033	983 292	1 016	982 784	38 176 347	38.83
44	0.001103	982 276	1 083	981 735	37 193 563	37.86
45	0.001181	981 193	1 159	980 614	36 211 828	36.91
46	0.001274	980 034	1 249	979 410	35 231 214	35.95
47	0.001389	978 785	1 360	978 105	34 251 804	34.99
48	0.001527	977 425	1 493	976 679	33 273 699	34.04
49	0.00169	975 932	1 649	975 108	32 297 020	33.09
50	0.001873	974 283	1 825	973 371	31 321 912	32.15
51	0.002074	972 458	2 017	971 450	30 348 541	31.21
52	0.002295	970 441	2 227	969 328	29 377 091	30.27

续表

中国人寿保险业经验生命表（2000—2003）（女）						
China Life Insurance Mortality Table（2000—2003）						
非养老金业务女表 CL2（2000—2003）						
年龄	死亡率	生存人数	死亡人数	生存人年数		平均余命
(x)	q_x	l_x	d_x	L_x	T_x	$\mathring{e}_x$
53	0.002546	968 214	2 465	966 982	28 407 763	29.34
54	0.002836	965 749	2 739	964 380	27 440 781	28.41
55	0.003178	963 010	3 060	961 480	26 476 401	27.49
56	0.003577	959 950	3 434	958 233	25 514 921	26.58
57	0.004036	956 516	3 860	954 586	24 556 688	25.67
58	0.004556	952 656	4 340	950 486	23 602 102	24.78
59	0.005133	948 316	4 868	945 882	22 651 616	23.89
60	0.005768	943 448	5 442	940 727	21 705 734	23.01
61	0.006465	938 006	6 064	934 974	20 765 007	22.14
62	0.007235	931 942	6 743	928 571	19 830 033	21.28
63	0.008094	925 199	7 489	921 455	18 901 462	20.43
64	0.009059	917 710	8 314	913 553	17 980 007	19.59
65	0.010148	909 396	9 229	904 782	17 066 454	18.77
66	0.011376	900 167	10 240	895 047	16 161 672	17.95
67	0.01276	889 927	11 355	884 250	15 266 625	17.15
68	0.014316	878 572	12 578	872 283	14 382 375	16.37
69	0.016066	865 994	13 913	859 038	13 510 092	15.6
70	0.018033	852 081	15 366	844 398	12 651 054	14.85
71	0.020241	836 715	16 936	828 247	11 806 656	14.11
72	0.022715	819 779	18 621	810 469	10 978 409	13.39
73	0.025479	801 158	20 413	790 952	10 167 940	12.69
74	0.028561	780 745	22 299	769 596	9 376 988	12.01
75	0.031989	758 446	24 262	746 315	8 607 392	11.35
76	0.035796	734 184	26 281	721 044	7 861 077	10.71
77	0.040026	707 903	28 335	693 736	7 140 033	10.09
78	0.044726	679 568	30 394	664 371	6 446 297	9.49
79	0.049954	649 174	32 429	632 960	5 781 926	8.91

续表

中国人寿保险业经验生命表（2000—2003）（女）						
China Life Insurance Mortality Table（2000—2003）						
非养老金业务女表 CL2（2000—2003）						
年龄	死亡率	生存人数	死亡人数	生存人年数		平均余命
(x)	q_x	l_x	d_x	L_x	T_x	$\overset{\circ}{e}_x$
80	0.055774	616 745	34 398	599 546	5 148 966	8.35
81	0.062253	582 347	36 253	564 221	4 549 420	7.81
82	0.069494	546 094	37 950	527 119	3 985 199	7.3
83	0.077511	508 144	39 387	488 451	3 458 080	6.81
84	0.086415	468 757	40 508	448 503	2 969 629	6.34
85	0.096294	428 249	41 238	407 630	2 521 126	5.89
86	0.107243	387 011	41 504	366 259	2 113 496	5.46
87	0.119364	345 507	41 241	324 887	1 747 237	5.06
88	0.132763	304 266	40 395	284 069	1 422 350	4.67
89	0.147553	263 871	38 935	244 404	1 138 281	4.31
90	0.16385	224 936	36 856	206 508	893 877	3.97
91	0.181775	188 080	34 188	170 986	687 369	3.65
92	0.201447	153 892	31 001	138 392	516 383	3.36
93	0.222987	122 891	27 403	109 190	377 991	3.08
94	0.246507	95 488	23 538	83 719	268 801	2.82
95	0.272115	71 950	19 579	62 161	185 082	2.57
96	0.299903	52 371	15 706	44 518	122 921	2.35
97	0.329942	36 665	12 097	30 617	78 403	2.14
98	0.362281	24 568	8 901	20 118	47 786	1.95
99	0.396933	15 667	6 219	12 558	27 668	1.77
100	0.433869	9 448	4 099	7 399	15 110	1.6
101	0.473008	5 349	2 530	4 084	7 711	1.44
102	0.514211	2 819	1 450	2 094	3 627	1.29
103	0.557269	1 369	763	988	1 533	1.12
104	0.601896	606	365	424	545	0.9
105	1	241	241	121	121	0.5

附录二　常用概率分布及其性质

一、离散分布

■ **泊松分布** Poisson(λ)，$\lambda>0$

- 概率分布：$P\{X=x\}=\dfrac{\lambda^x}{x!}e^{-\lambda}$，$x=0$，1，2，…
- 均值、方差：$E(X)=Var(X)=\lambda$，
- 矩母函数、母函数：

$$M_X(t)=E(e^{tX})=e^{\lambda(e^t-1)}$$

$$P_X(t)=E(t^X)=e^{\lambda(t-1)}$$

- 性质：当贝努利概型中的试验次数（n）很大，而每次成功的概率（p）很小时，试验成功次数接近于常数情况下的泊松分布。

■ **负二项分布** NB(k，p)，$k>1$　$0<p<1$

- 概率分布：$P\{X=x\}=\begin{pmatrix}k+x-1\\x\end{pmatrix}p^kq^x$，　$x=0$，1，2，…
- 均值、方差：$E(X)=\dfrac{kq}{p}$，　$Var(X)=\dfrac{kq}{p^2}$
- 矩母函数、母函数：

$$M_X(t)=E(e^{tX})=\left(\frac{p}{1-qe^t}\right)^k,\quad t<-\ln q$$

$$P_X(t)=E(t^X)=\left(1-\frac{q}{p}(t-1)\right)^{-k}$$

- 性质：贝努利试验系列中第 k 次成功正好出现在第 $x+k$ 次试验上的概率。x 为第 k 次成功前失败的次数。

■ **几何分布** Geo(p)，$0<p<1$

- 概率分布：$P(X=x)=pq^x$，　$x=0$，1，2，…；$p+q=1$
- 均值、方差：$E(X)=\dfrac{q}{p}$，　$Var(X)=\dfrac{q}{p^2}$
- 矩母函数、母函数：

$$M_X(t)=E(e^{tX})=\frac{p}{1-qe^t},\quad t<-\ln q$$

$$P_X(t)=E(t^X)=\left(1-\frac{q}{p}(t-1)\right)^{-1}$$

- 性质：

（1）n 重贝努利实验中首次成功正好出现在第 $x+1$ 次实验上的概率。p 为单次成功概率，x 是首次成功前试验失败的次数。

（2）几何分布是负二项分布当参数 $k=1$ 时的特殊情形。

■ **二项分布** $B(n, p)$，$0<p<1$，$n=0, 1, 2, \cdots$

- 概率分布：$P\{X=x\}=\binom{n}{x}p^x q^{n-x}$，$x=0, 1, 2, \cdots, n$；$p+q=1$
- 均值、方差：$E(X)=np$，$Var(X)=npq$
- 矩母函数、母函数：

  $M_X(t)=E(e^{tX})=(q+pe^t)^n$，$-\infty<t<+\infty$

  $P_X(t)=E(t^X)=(1+p(t-1))^n$
- 性质：贝努利实验中正好成功 x 次的概率，其中 p 为每次成功概率。

二、连续分布

■ **均匀分布** $U(a, b)$

- 密度函数：$f_X(x)=\dfrac{1}{b-a}$，$a<x<b$
- 分布函数：$F_X(x)=\dfrac{x-a}{b-a}$，$a<x<b$；$F(x)=0$，$x\leqslant a$，$F(x)=1$，$x\geqslant b$
- 均值、方差、r 阶原点矩：

$$E(X)=\frac{1}{2}(a+b)$$

$$Var(X)=\frac{1}{12}(b-a)^2$$

$$E(X^r)=\frac{b^{r+1}-a^{r+1}}{(r+1)(b-a)}$$

■ **指数分布** $Exp(\theta)$，$\theta>0$

- 密度函数：$f_X(x)=\theta^{-1}e^{-x/\theta}$，$x>0$
- 分布函数：$F_X(x)=1-e^{-x/\theta}$，$x>0$
- 危险率函数 $h(x)=\dfrac{1}{\theta}$
- 均值、方差、r 阶原点矩：

  $E(X)=\theta$

  $Var(X)=\theta^2$

  $E(X^r)=\theta^r\Gamma(r+1)$
- 矩母函数：$M_X(t)=(1-t\theta)^{-1}$，$t<\lambda$
- 性质：

（1）指数分布又称寿命分布，它是任何风险单位寿命分布的近似。

（2）指数分布具有“无记忆”的特点，即若 $X\sim Exp(\lambda)$，则 $\forall a>0$，

有 $X-a|X>a \sim Exp(\lambda)$。

■ **伽玛分布** Gamma$(\alpha,\ \theta)$，$\alpha>0$，$\lambda>0$

- 密度函数 $f_X(x)=\dfrac{1}{\theta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\theta}$，$x>0$
- 分布函数 $F_X(x)=\Gamma(\alpha;\ x/\theta)=\dfrac{1}{\Gamma(\alpha)}\int_0^{x/\theta}t^{\alpha-1}e^{-t}dt$，$\alpha>0$，$x>0$

其中，$\Gamma(\alpha)=\int_0^{\infty}t^{\alpha-1}e^{-t}dt$

- 均值、方差、r 阶原点矩：

$E(X)=\alpha\theta$

$Var(X)=\alpha\theta^2$

$E(X^r)=\dfrac{\theta^r\Gamma(r+\alpha)}{\Gamma(\alpha)}$，　$r>-\alpha$

$E(X^k)=\theta^k(\alpha+k-1)\cdots\alpha$，当 k 为正整数

- 矩母函数 $M_X(t)=(1-t\theta)^{-\alpha}$，$t<\lambda$
- 性质：

（1）若 α 较小，伽玛分布的密度函数偏度较大，并且向右逐渐减小；若 α 较大，则密度函数比较对称，且可按正态分布近似。

（2）$\alpha=1$ 时，退化为指数分布。

（3）若 $X\sim$ Gamma$(\alpha,\ \lambda)$，则 $2\lambda X\sim$ Gamma$\left(\alpha,\ \dfrac{1}{2}\right)$（这是 $\chi^2_{2\alpha}$ 分布）

（4）如果 $X_1,\cdots,X_n$ 是分别服从参数为 $(\alpha_i,\ \theta)$ 的伽玛分布的独立随机变量则 $Y_1=X_1+\cdots+X_n$ 服从参数为 $\left(\sum_{i=1}^{n}\alpha_i,\ \theta\right)$ 的伽玛分布。

■ **对数正态分布** $LogN(\mu,\ \sigma^2)$，$-\infty<\mu<+\infty$，$\sigma>0$

- 密度函数：$f_X(x)=\dfrac{1}{\sqrt{2\pi}\sigma}\dfrac{e^{-\frac{z^2}{2}}}{x}$，$z=\left(\dfrac{\ln x-\mu}{\sigma}\right)$，$x>0$
- 分布函数：$F(x)=\Phi(z)$，$\Phi(\cdot)$ 为标准正态分布的分布函数
- 均值、方差、r 阶原点矩：

$E(X)=e^{\mu+\frac{1}{2}\sigma^2}$

$Var(X)=e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$

$E(X^r)=e^{\mu r+\frac{1}{2}\sigma^2r^2}$

- 性质：

（1）如果 $X\sim LogN(\mu,\ \sigma^2)$，则 $LogX\sim N(\mu,\ \sigma^2)$；

（2）设 a 和 b 为正实数，X 服从参数为 μ 和 σ^2 的对数正态分布，则 $Y=aX^b$ 仍服从对数正态分布，其参数为 $b\mu+\ln a$ 和 $b^2\sigma^2$。

■ **帕累托分布** Pareto$(\alpha,\ \lambda)$，$\alpha>0$，$\lambda>0$

- 密度函数：$f_X(x) = \frac{\alpha\lambda^{\alpha}}{(\lambda + x)^{\alpha+1}}$，$x > 0$
- 分布函数：$F_X(x) = 1 - \left(\frac{\lambda}{\lambda + x}\right)^{\alpha}$
- 均值、方差、r 阶原点矩：

$$E(X) = \frac{\lambda}{\alpha - 1}, \ (\alpha > 1)$$

$$Var(X) = \frac{\alpha\lambda^2}{(\alpha - 1)^2(\alpha - 2)}, \ \alpha > 2;$$

$$E(X^r) = \frac{\lambda^r\Gamma(r+1)\Gamma(\alpha - r)}{\Gamma(\alpha)}, \ \alpha > r;$$

$$E(X \wedge x) = \frac{\lambda}{\alpha - 1}\left[1 - \left(\frac{\lambda}{x + \lambda}\right)^{\alpha - 1}\right], \ \alpha \neq 1$$

$$E(X \wedge x) = -\lambda \log\left(\frac{\lambda}{x + \lambda}\right), \ \alpha = 1$$

- 性质：

（1）当 $\alpha \to \infty$ 时，$Var(X)/[E(X)]^2$ 趋于 1。

（2）当均值 $\mu = E(X)$ 保持不变，令 $\alpha \to \infty$，则该帕累托分布收敛到指数分布。

■ **韦伯分布** Weibull(γ, θ)，$c > 0$，$\gamma > 0$

- 密度函数：$f(x) = \frac{\gamma}{\theta}x^{\gamma - 1}e^{-\frac{1}{\theta}x^{\gamma}}$ $x > 0$
- 分布函数：$F_X(x) = 1 - e^{-x^{\gamma}/\theta}$
- 危险率函数 $h(x) = \frac{\gamma}{\theta}x^{\gamma - 1}$，$x \geqslant 0$，$c > 0$，$\gamma > 0$
- 均值、方差、r 阶原点矩：

$$E(X) = \Gamma(1 + 1/\gamma)\theta^{1/\gamma}$$

$$Var(X) = \Gamma(1 + 2/\gamma)\theta^{2/\gamma} - [\Gamma(1 + 1/\gamma)\theta^{1/\gamma}]^2$$

$$E(X^r) = \theta^{\frac{r}{\gamma}}\Gamma\left(1 + \frac{r}{\gamma}\right)$$

- 性质：对于指数分布，用 x^{γ} 代替 x 则得韦伯分布。

■ **贝塔分布** $\beta(\alpha, \beta)$，$\alpha > 0$，$\beta > 0$

- 密度函数：$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha - 1}(1 - x)^{\beta - 1}$，$0 < x < 1$
- 分布函数：$F_X(x) = \beta(\alpha, \beta; x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^x u^{\alpha - 1}(1 - u)^{\beta - 1}du$
- 均值、方差、r 阶原点矩：

$$E(X) = \frac{\alpha}{\alpha + \beta}$$

$$Var(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$E(X^r)=\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+r)}{\Gamma(\alpha+\beta+r)\Gamma(\alpha)}$$

■ **广义帕累托分布** Pareto(α, λ, k), $\alpha>0$

- 密度函数：$f_X(x)=\dfrac{\Gamma(\alpha+k)\lambda^\alpha x^{k-1}}{\Gamma(\alpha)\Gamma(k)(\lambda+x)^{\alpha+k}}$, $x>0$
- 分布函数：$F_X(x)=\beta(k,\alpha;u)$, $u=\dfrac{x}{x+\lambda}$
- 均值、方差、r 阶原点矩：

$$E(X)=\frac{\lambda k}{\alpha-1},\ (\alpha>1);$$

$$Var(X)=\frac{\lambda^2k(k+\alpha-1)}{(\alpha-1)^2(\alpha-2)},\ \alpha>2;$$

$$E(X^r)=\frac{\lambda^r\Gamma(r+k)\Gamma(\alpha-r)}{\Gamma(\alpha)\Gamma(k)},\ \alpha>r$$

■ **布尔分布** Burr(α, λ, γ) $\alpha>0$, $\gamma>0$, $\lambda>0$

- 密度函数：$f_X(x)=\dfrac{\alpha\gamma\lambda^\alpha x^{\gamma-1}}{(\lambda+x^\gamma)^{\alpha+1}}$, $x>0$
- 分布函数：$F_X(x)=1-\left(\dfrac{\lambda}{\lambda+x^\gamma}\right)^\alpha$
- 均值、方差、r 阶原点矩、矩母函数：

$$E(X)=\frac{\lambda^{\frac{1}{\lambda}}\Gamma\left(\alpha-\frac{1}{\gamma}\right)\Gamma\left(1+\frac{1}{\gamma}\right)}{\Gamma(\alpha)},\ \alpha>\frac{1}{\gamma};$$

$$Var(X)=\frac{\lambda^{\frac{2}{\lambda}}\Gamma\left(\alpha-\frac{2}{\gamma}\right)\Gamma\left(1+\frac{2}{\gamma}\right)}{\Gamma(\alpha)}-[E(X)]^2,\ \alpha>\frac{2}{\gamma};$$

$$E(X^r)=\frac{\lambda^{\frac{r}{\lambda}}\Gamma\left(\alpha-\frac{r}{\gamma}\right)\Gamma\left(1+\frac{r}{\gamma}\right)}{\Gamma(\alpha)},\ \alpha>\frac{r}{\gamma};$$

- 注释：帕累托分布的另一推广形式，用 x^r 替代帕累托分布中的 x。

■ **Gompertz 分布**

- 密度函数 $f(x)=Bc^xe^{\frac{B}{\ln c}(1-c^x)}$
- 生存函数 $S(x)=e^{\frac{B}{\ln c}(1-c^x)}$　$x\geqslant0$, $B>0$, $c>1$
- 危险率函数 $h(x)=Bc^x$,　$x\geqslant0$, $B>0$, $c>1$
- 注释：常用于生存分析中，分布的期望也不易求得。

■ **Makeham 分布**

- 生存函数 $S(x)=e^{\frac{B}{\ln c}(1-c^x)-Ax}$

- 危险率函数 $h(x)=A+Bc^x$，$x\geqslant 0$，$B>0$，$c>1$，$A>-B$
- 注释：常用于生存分析中，分布的期望也不易求得。

■ **正态分布** $N(\mu,\sigma)^2$，$-\infty<\mu<+\infty$，$\sigma>0$

- 密度函数：$f_X(x)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$，$(-\infty<x<+\infty)$
- 均值、方差：

  $E(X)=\mu$

  $Var(X)=\sigma^2$

- 矩母函数：$M_X(t)=E(e^{tX})=e^{\mu t+\frac{\sigma^2 t^2}{2}}$
- 性质：设 X_1，$X_2,\cdots,X_n$ 是独立同分布的随机变量，满足 $E(X)<\infty$，$E(X^2)<\infty$，$S=\sum_{i=1}^{n}X_i$，则$\dfrac{S-E(S)}{\sqrt{Var(S)}}$依分布收敛到 $N(0,1)$。

■ **χ^2 分布**（χ_n^2，$n=1,2,3,\cdots,n$ 自由度）

- 密度函数：$f_X(x)=\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}$，$x>0$
- 均值、方差、r 阶原点矩：

  $E(X)=n$

  $Var(X)=2n$

  $$E(X^r)=\frac{2^r\Gamma\left(r+\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

- 矩母函数：$M_X(t)=E(e^{tX})=(1-2t)^{-\frac{n}{2}}$，$t<\dfrac{1}{2}$
- 性质：（1）n 个独立的标准正态分布随机变量的平方和服从χ^2 分布（2）χ^2 分布是伽马分布的特例，$\chi_n^2=Gamma\left(\frac{n}{2},\frac{1}{2}\right)$。$\chi^2$ 分布常用来检验对分布的拟合是否恰当。

■ **t 分布** t_n，$n=1,2,3,\cdots$，自由度 n

- 密度函数：$f_X(x)=\dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\left(1+\dfrac{x^2}{n}\right)^{-\frac{n+1}{2}}$，$-\infty<x<+\infty$
- 均值、方差：

  $E(X)=0$，　　$n>1$

  $Var(X)=\dfrac{n}{n-2}$，　　$n>2$

- 性质：如果 $X_1 \sim N(0,1)$ 和 $X_2 \sim \chi_n^2$ 相互独立，则 $\frac{X_1}{\sqrt{X_2/n}} \sim t_n$

■ **F 分布** $F_{m,n}$，$m=1, 2, 3, \cdots, n=1, 2, 3, \cdots$

- 密度函数：$f_X(x) = \left(\frac{m}{n}\right)^{\frac{n}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} x^{\frac{m}{2}-1}\left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}}$，$x>0$

- 均值、方差、r 阶原点矩：

$$E(X) = \frac{n}{n-2}, \quad n>2$$

$$Var(X) = \frac{2n^2(m+n-2)}{m(n-4)(n-2)^2}, \quad n>4$$

$$E(X^r) = \frac{\left(\frac{n}{m}\right)^r \Gamma\left(\frac{m}{2}+r\right)\Gamma\left(\frac{n}{2}-r\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}, \quad n>2r$$

- 性质：

（1）如果 $X_1 \sim \chi_m^2$ 和 $X_2 \sim \chi_n^2$ 相互独立，则 $\frac{X_1/m}{X_2/n} \sim F_{m,n}$

（2）$F_{m,n}(1-\alpha) = \frac{1}{F_{n,m}(\alpha)}$

（3）F 分布用于检验对方差的估计。

附录三　部分习题解答

第二章

1. （1） $f(36)=0.00625$；（2）$\lambda(50)=0.01$；

   （3）$\Lambda(75)=0.6931$；（4）$E(T)=66.7$；

   （5）$E(T^2)=\frac{16\ 000}{3}$；$Var(T)=888.9$

2. $45\sqrt{2}$

4. $F(75)=0.9375$；$f(75)=0.02$；$\mu(75)=0.08$

5. （1） 不变；（2） 递增；（3） 递减。

6. ${}_2m_3=1/4$

7. 该试验设备未来寿命的中位数是 4.7504

8. $S(x)=1-(0.01x)^2=0.5$，解得 $x=70.71067$，$\therefore e_{70.71067}=15.4822$

11. $q'^{(1)}_{67}=1-\left(1-\frac{4}{19.998}\right)^{\frac{1}{4}}=0.05426$

12. （1）${}_tp_x^{(\tau)}={}_tp'^{(1)}_x\,{}_tp'^{(2)}_x=\begin{cases}1-0.15t, & t<0.5\\ 0.97(1-0.15t), & t\geqslant 0.5\end{cases}$

    （2） $q_x^{(1)}=0.1478$；$q_x^{(2)}=0.02775$

13. 在 60 岁之前退休的概率为 0.0689。

14. ${}_2q_x^{(2)}=0.644$

15. $Var(k\wedge 3)=1.07$

16. $h(4)=1.202553$

17. $q_x^{(2)}=0.259$

18. $p_x^{\prime(2)}=0.512$

19. $q_x^{(1)}=0.1802$

20. $d_{60}^{(3)}+d_{61}^{(3)}=265.63$

第三章

1. （1） $S(0)=1$，$S(1)=0.9$，$S(2)={}_2p_0S(0)=0.72$，$S(3)={}_3p_0S(0)$ $=0.432$，$S(4)={}_4p_0S(0)=0.1296$，$S(5)=0$

   （2） $l_0=10\ 000$，$l_1=9\ 000$，$l_2=7\ 200$，$l_3=4\ 320$，$l_4=1\ 296$，$l_5=0$

   $d_0=1\ 000$，$d_1=1\ 800$，$d_2=2\ 880$，$d_3=3\ 024$，$d_4=1\ 296$

(3) $\omega=5$

2. (1) ${}_3d_0=5\ 680$；(2) ${}_2q_1=0.2$；(3) ${}_3p_1=0.144$；(4) ${}_3q_2=1$

3. ${}_{x|}q_0\ \dfrac{s(0)-s(x)}{s(0)}$，${}_{x|}q_0=\dfrac{l_x-l_{x+1}}{10}$

4. (1) $w=90$；(2) ${}_{60}p_0=0.2$；(3) ${}_{20|15}q_{10}=0.2083$

5. ${}_3d_1=2\ 081.61$

6. $\dfrac{\partial}{\partial x}{}_sp_x=\dfrac{\partial}{\partial x}\left(\dfrac{S(x+s)}{S(x)}\right)=\dfrac{S'(x)S(x+s)-S'(x+s)S(x)}{S^2(x)}$

$=-{}_sp_x\mu_x-\dfrac{S'(x+s)}{S(x)}$

7. ${}_{10}m_{70}=\dfrac{2}{15}$

9. (1) UDD：$t=25.1$；(2) 常值死力：25.0968195

10. $l_{[0]}=150\ 000$

11. (1) 在死亡均匀分布下，${}_5q_{30}=\dfrac{332.5}{98\ 617}=0.0033716$

(2) 设 μ 是恒定的死力，${}_5q_{30}=1-{}_{10}p_{30}=1-\exp(-5\mu)=0.0033773$

(3) 均匀分布下活过 35 岁的人数为：$98\ 617(1-0.0033716)=98\ 284.5$

在死力恒定下活过 35 岁的人数为：$98\ 617(1-0.0033773)=98\ 283.9$

13. (1) 死亡均匀分布时，$m_x=0.105263$

(2) 死力恒定时，$m_x=0.105361$

(3) 在 Balducci 假设条件下，$m_x=0.1054580$

14. (1) $l_0=1\ 000$；$l_{120}=0$；$d_{33}=8.33$；${}_{20}p_{30}=0.778$；${}_{30}q_{20}=0.3$

(2) ${}_{20|5}q_{25}=0.0526$；(3) $p=0.0746$

15. $q_{x+1}+q_{x+2}=0.1+0.05=0.15$

16. 在巴尔杜奇假设下，$l_{65.5}=88.89$

17. $e_{[81]}=\dfrac{8\ 000+460-910}{920}=8.21$

18. ${}_{4|14}q_{50}=0.3783$

19. ${}_{1.5}q_{x+0.25}=0.82$

第四章

1. $E(X\wedge 5\ 000)=1\ 038.26$

2. $P[S<90|S>65]=0.9206$

3. $c=a=\beta/(1+\beta)=0.29$

4. $P(t>12)=36.8\%$

5. $F_X(200)=0.7566$

6. 0.16

7. $E(Y \text{ per payment}) = 38.91$

8. 没有任何支付的概率为$f(0) = e^{-1.4583} = 0.23$

9. 总的等待时间为26.65。

10. 9.43%

11. 329

12. $\alpha = 0.42$

13. 2 700

14. $p(N=2) = p \times 0.25 + (1-p) \times 0.375 = 0.375 - 0.125p$

15. $E(Y) = 38.52$

16. $P(N=0) = \frac{1}{2}$

17. $EN = 2.75$, $VarN = 2.0625$

18. N^*服从负二项分布，参数为$r = 10$，$\theta^* = 0.15$。

19. $c = 0.44$

20. $p_k = \left(a + \frac{b}{k}\right)p_{k-1}$

$p_1 = 0.25 = (a+b) \times p_1 = (a+b) \times 0.25 \Rightarrow a + b = 1$

$p_2 = 0.1875 = \left(a + \frac{b}{2}\right) \times p_1 = \left(a + \frac{b}{2}\right) \times 0.25 \Rightarrow \left(1 - \frac{b}{2}\right) \times 0.25$

$= 0.1875$

$a = 0.5$, $b = 0.5$

$p_3 = 0.125$

第五章

1. $E(Y) = 3.5 \times 2.5 = 8.75$

$Var(Y) = 22.6$

2. $E(S) = 0.02 \times \frac{a}{2} = 0.01a$; $Var(Y) = 0.00657a^2$

3. $f_S(1) = 1$

4. $E(S) = \frac{3}{2}$, $Var(S) = \frac{1}{4}$, $P(S \leqslant 0.5) = \frac{1}{48}$, $P(S \leqslant 10) = \frac{1}{6}$

5. $S \sim \begin{pmatrix} 0 & 100 & 200 & 300 & 400 & 500 & 600 \\ 0.512 & 0.288 & 0.15 & 0.39375 & 0.009375 & 0.001125 & 0.000125 \end{pmatrix}$

$E(S) = E(X_1 + X_2 + X_3) = 75$

6. $E(S^3) = 10\ 626$

7. (1) $E(U) = 18$; $Var(U) = 36$

(2) $U \sim N(18, 6^2)$

$y_{0.05}=P(U>y_{0.05})=0.05\Rightarrow P(U<y_{0.05})=0.95\Rightarrow\Phi\left(\frac{y_{0.05}-18}{6}\right)=0.95$

$\Rightarrow\frac{y_{0.05}-18}{6}=\Phi^{-1}(0.95)\Rightarrow y_{0.05}=6\Phi^{-1}(0.95)+18=29.28$

$y_{0.1}=6\Phi^{-1}(0.9)+18=37.15$

(3)推测 $U\sim Gamma(\alpha,\beta)$，$\frac{\alpha}{\beta}=18$，$\frac{\alpha}{\beta^2}=36\Rightarrow\alpha=9,\ \beta=0.5$

可得到：$y_{0.05}=29.28$，$y_{0.1}=37.15$

8. $c=1$，$b=-1$，$d=a$，$E(S)=1.5$，$Var(S)=0.25$，

$P(S\leqslant 0.5)=\Phi\left(\frac{0.5-1.5}{0.5}\right)=1-\Phi(2)=0.02275$

$P(S<1)=\Phi\left(\frac{1-1.5}{0.5}\right)=1-\Phi(1)=0.1587$

9. (1) $E(N)=6.4$，$Var(N)=6.144$；(2) $E(S)=70\ 000$；$Var(S)=17.072\times10^8$；

(3) $\theta=1.373$

10. $E(S_R)=172$，$Var(S_R)=16.07^2$，$\theta=0.154$

第六章

2. (1) $M_N(t)=\frac{e^{(n+1)t}-1}{(n+1)(e^t-1)}$；(2) $E(N)=M_N'(0)=\frac{n}{2}$

(3) $Var(N)=M_N''(0)-[M_N'(0)]^2=\frac{n}{6}+\frac{n^2}{12}$

4. $E(S)=6\times3.4=20.4$；$Var(S)=72$

5. $E(S)=90$；$Var(S)=10\ 908.1$

6. $E(X)=1.6$；$Var(X)=0.44$

7. $f_s(3)=0.04736$

8. (1) $M_S(t)=\left(\frac{t}{1-2t}\right)^3$；(2) $E(S)=0$，$Var(S)=0$

9. $P(1)=0.4$，$P(2)=0.4$，$P(3)=0.2$

10. $P(S\leqslant 2)=0.0299$

11. $\lambda_1=10$，$\lambda_2=11$，$\lambda_3=8$

12. $f(0)=0.4^8$，$f(1)=1.2f(0)$，$f(2)=2.61f(0)$，$f(3)=4.635f(0)$
$f(4)=1.6875f(0)$，$f(5)=7.624f(0)$，$f(6)=8.4363f(0)$

14. 见表 6－14

15. (1) $S\sim Gamma(X-d,\ n,\ \beta)$

(2) $E(S)=d+\frac{\lambda}{\beta}$，$Var(S)=\frac{\alpha\lambda}{\beta}$

16. $M_{X_1}(t)=e^{t\times0}\times0.5+e^{t\times100}\times0.3+e^{t\times400}\times0.1+e^{t\times900}\times0.1$

$M_{X_2}(t)=e^{t\times 0}\times 0.25+e^{t\times 50}\times 0.5+e^{t\times 350}\times 0.15+e^{t\times 850}\times 0.1$

$M_{S_1}(t)=M_{N_1}(\log M_{X_1}(t))=e^{0.3(e^{t\times 0}\times 0.5+e^{t\times 100}\times 0.3+e^{t\times 400}\times 0.1+e^{t\times 900}\times 0.1)}$

$M_{S_2}(t)=M_{N_1}(\log M_{X_2}(t))=e^{0.3(e^{t\times 0}\times 0.25+e^{t\times 50}\times 0.5+e^{t\times 350}\times 0.15+e^{t\times 850}\times 0.1)}$

$M_S(t)=M_{S_1}(t)M_{S_2}(t)$

$E(S)=M_S'(t)=161$

$Var(S)=M_S''(t)-[M_S'(t)]^2=70\ 829$

17. $1-\Phi(0.123)$

18. $v=\sigma^3\sqrt{\frac{2\sigma}{\mu}}$

19. 正态近似 $P(S\leqslant 8)=0.9865$；平移伽玛近似 $P(S\leqslant 8)=0.9619$

20. 0.180246，0.193154，0.169722

21. $Var(S)=Var(S_1)+Var(S_2)=24.6625\approx 25$

第七章

4. $u=\frac{4}{\lambda-4}\ln\frac{0.2}{\lambda}$

6. $\theta=\frac{65}{31}$

7. $\Psi(u)=\frac{24}{35}e^{-u}+\frac{1}{35}e^{-6u},\ u\geqslant 0$

8. $\theta=\frac{2}{3}$，$R=2$

9. 0.404

11. $E(L)=\frac{p_2}{2\theta p_1}$，$Vae(L)=\frac{p_3}{3\theta p_1}+\frac{1}{4}\left(\frac{p_2}{\theta p_1}\right)^2$

12. $\theta=\frac{1}{4}$，$\Psi(0)=\frac{4}{5}$，$E(L_1)=\frac{5}{2}$，$Var(L_1)=\frac{25}{12}$，$E(N)=4$，

    $Var(N)=20$，$E(L)=10$，$Var(L)=\frac{400}{3}$

13. $\Psi(2,2)=0.96$

15. 17 400 000

17. $u=43.75$

18. $r=6$

19. 破产的概率为 $\Psi(1,2)=1-0.81=0.19$

20. (1) $E(L)=39/76=0.51$

    (2) $\psi(u)=0.068$

第八章

1. $\hat{S}(6.5)=0.48$

2. 0.00533

3. $[0.475, 0.994]$
4. $|S^T(5) - S^B(5)| = 0.08472$
5. $[1.665, 1.967]$
6. $\hat{H}(300) = 0.9163$
7. 0.09
8. 0.03072
9. $S_{10}(1.6) = 0.7143$
10. $\hat{S}(t_9) = 0.7$
11. $\hat{S}(t_0) = 0.7386$
12. $n = 13$
13. $\hat{H}(t) = 0.5456$
14. $\hat{S}(t) = 0.6497$
15. $|\hat{H}_2(75) - \hat{H}_1(75)| = 0.112$
16. $[[0.1760, 1.3576]]$
17. 5
18. 100
19. $|\hat{H}_1(7\,000) - \hat{H}_2(7\,000)| = 0.5833$
20. 39
21. 0.3
22. 0.53125
23. $[1, 2]$
24. 0.485
25. 774

第九章

1. $\hat{\alpha} = 7.632135$, $\hat{\theta} = 497.8947$
2. $\hat{\alpha} = 2.442$, $\hat{\theta} = 2\,053.985$
3. $\hat{p} = 0.856$
4. $\hat{p} = 0.856$
5. (1) $\hat{\theta} = 17.06$; (2) $\hat{\theta} = 33.33$
6. $\hat{\theta} = 25$
7. (1) $\hat{\theta} = 30$; (2) $\hat{\theta} = 12.5$; (3) $\hat{\theta} = 7.5$
8. $\hat{\theta} = \dfrac{312}{8}$
9. $\hat{\theta} = 75$

10. （1）$\hat{\theta}=14$；（2）$\hat{\theta}=40$

11. $\hat{\alpha}=1.26$，$P(X\leqslant 10)$的极大似然估计为0.582。

12. $\hat{\theta}=13.8$

13. $\hat{\theta}=14$

14. $\hat{\theta}=11.85$，$P(X\leqslant 10)=0.609$

15. $\hat{\theta}=26.5$，参数θ的置信度95%的置信区间为（11.5，41.5）。

16. $Var(\hat{\theta}^2)=164\ 385$

17. $P[X>10]$的置信度为95%的置信区间为（0.539，0.832）。

18. $\widehat{Var}[\hat{\theta}]=\frac{\hat{\theta}^2}{n}=\frac{\bar{X}^2}{n}=\frac{26.5^2}{12}=58.5$

19. 置信度为95%的置信区间为：

$\hat{\theta}^2\pm 1.96\sqrt{\widehat{Var}[\hat{\theta}^2]}=702.25\pm 1.96\sqrt{164\ 327}=(-92.3,1\ 496.8)$

第十章

1. 检验统计量落在97.5%置信水平临界值（7.38）和99%置信水平临界值（9.21）之间，在前者置信度下拒绝原假设，在后者置信度下无法拒绝。

2. $\hat{\theta}=8\ 550/1\ 125=7.6$

3. 0.285

4. 模型在2.5%显著性水平下拒绝原假设，但在1%显著性水平下无法拒绝。

5. 6.65

6. 样本量趋于正无穷时，只有Kolmogorov－Smirnov检验统计量会趋于0，因为样本量的大小出现在在其临界值的表达式的分母上，另外两个统计量将趋于无穷，因为而样本量的大小出现在Ånderson－Darling检验统计量和χ^2拟合优度检验统计量的分子上。

7. 0.402

8. 在2.5%显著性水平下拒绝原假设，但在1%显著性水平下无法拒绝。

9. 在5%显著性水平下拒绝原假设，2.5%显著性水平下无法拒绝原假设。

10. -0.071

11. 最大离差为0.195。

12. Kolmogorov－Smirnov检验统计量为0.136．无法拒绝原假设．Ånderson－Darling检验统计量为0.3032的结果。

13. （1）检验统计量为9.8，在5%显著水平下拒绝原假设；（2）合并

0. 1 或合并 1. 2 或合并 3. 4 都满足要求。

14. 统计量为 7. 0，p 值为 0. 032。

15. $l_c > -157.48$

16. （1） D = 0. 3132；（2） 1. 483

17. Q = 7. 2

18. B 最多可以有 4 个参数。

第十一章

2. （1） 0. 002； （2） 19. 951；1 491. 511； （3） $a = 1.337$，$b = -0.0077$

3. （1） 38；（2） 3；（3） 13；（4） 14；（5） $x = 26, 27, 28, \cdots, 53$

4. 乙的方法具有较小方差。

5. $a_0 = \frac{11}{31}$，$v_{44} = \frac{10}{31}u_{43} + \frac{11}{31}u_{44} + \frac{10}{31}u_{45}$

6. $a = \frac{5}{3}$，$b = -\frac{1}{3}$

7. （1） 它们相同；

（2） 它们相同，M 在第二种情形下的极小值记为 M^*，其中$\frac{1}{w}$乘上 M 在第一种情形的值，但产生这个极小值的 V_x 是相同的。

8. 2

9. 1. 25

10. $V' = \left(\frac{13}{9}, \frac{74}{7}, 15\right)$

11. 0. 093

12. 15

13. （1） k；（2） 都是下降的；（3）$a_r = \dfrac{\sum_x n_{[x]+r}v_{x+r} - \sum_x \theta_{[x]+r}}{\sum_x n_{[x]+r}(v_{x+r} - v_{[x]+r})}$

14. （1） $c = 1.079$；$g = 0.997$；（2） 1. 07754

15. $n = 4.7681254$，$k = 5.8859 \times 10^{-11}$

16. $\hat{b} = \left[\frac{1}{2n}\sum_{i=1}^{n} y_i^2 - \frac{1}{2}x^2\right]^{-1}$

17. $\hat{b} = 18$

18. $p_1(x) = c_1 + c_2x + c_3x^2 + c_4(x - 29.7)^2$

$p_2(x) = c_1 + c_2x + c_3x^2 + c_4(x - 29.7)^2 + c_5(x - 62.5)^2$

$p_3(x) = c_1 + c_2x + c_3x^2 + c_4(x - 29.7)^2 + c_5(x - 62.5)^2 + c_6(x - 80.6)^2$

$$SS = \sum_{20}^{29} w_x(u_x - c_1 - c_2x - c_3x^2)^2 + \sum_{30}^{62} w_x[u_x - c_1 - c_2x - c_3x^2 - c_4(x - 29.7)^2]^2$$
$$+ \sum_{63}^{80} w_x[u_x - c_1 - c_2x - c_3x^2 - c_4(x - 29.7)^2 - c_5(x - 62.5)^2]^2$$
$$+ \sum_{63}^{80} w_x[u_x - c_1 - c_2x - c_3x^2 - c_4(x - 29.7)^2 - c_5(x - 62.5)^2 - c_6(x - 80.6)^2]^2$$

19. $F(0)u_x = 7A(0) + 5B(0)$

20. $V_{x+s} = [A(s) + A(t)]\mu_x + A(s)\Delta\mu_x + [B(s) + B(t)]\Delta^2\mu_{x-1} + B(s)\Delta^3\mu_{x-1} + [C(s) + C(t)]\Delta^4\mu_{x-2} + C(s)\Delta^5\mu_{x-1} + \cdots$

21. （1） 是密切的；（2） 是再生的。

第十二章

1. k = 0.0596
2. k = 56.25
3. r = 0.0596
4. q = 0.1
5. 4 328
6. k = 1.90
7. 0.5572
8. 3
9. 15
10. 0.68
11. 29/70
12. 0.148
13. r = 3
14. 3.27
15. 12 522.65
16. 2 594.58
17. 这两个风险的 Bühlmann 信度保费分别为：$\hat{Z}\bar{X}_1 + (1-\hat{Z})\hat{\mu} = 133/24$，$\hat{Z}\bar{X}_2 + (1-\hat{Z})\hat{\mu} = 203/24$
18. 48 000
19. 16.6
20. 92.64
21. 100.83

22. 第一组：$\hat{Z}_1\hat{X}_1+(1-\hat{Z}_1)\hat{\mu}=\frac{205}{216}\times 7+\frac{11}{216}\times 10=\frac{515}{72}$

第二组：$\hat{Z}_2\hat{X}_2+(1-\hat{Z}_2)\hat{\mu}=\frac{205}{216}\times 13+\frac{11}{216}\times 10=\frac{925}{72}$

23. 8.42

24. 10

25. 8

第十三章

1. 1，3，4
2. 泊松分布数值为 1，6，4
3. 到第三次成功的试验次数为 6
4. 0.6，1，1.6
5. 2.06
6. $n=\left(\frac{1.645}{0.05}\right)^2\left(\frac{s}{\bar{X}}\right)^2$，其中 $s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n-1}$，$\bar{X}=\frac{\sum_{i=1}^{n}x_i}{n}$
7. $n\geqslant 1\ 000$
8. $n=\frac{1\ 082.41P_n}{1\ 082.41-P_n}$
9. 最小样本量为 44。
10. 在时刻 2，索赔事件 1 和 2 处理完毕，索赔事件 3 正在处理中。
11. 保险公司第一年赔付 1 000；第二年赔付 21 000。
12. 总赔付额为 755.02。
13. 这四个月的总赔付为 1 274.02。
14. 保险公司的总赔付额为 2 901.28。
15. 保险公司在时刻 1.3879 破产。
16. $c>21.7$
17. (1)$MSE(x_e)=\frac{1}{k}\sum_{j=1}^{k}[\bar{x}(j)-\bar{x}]^2=\frac{26}{9}$；(2)$MSE(\max(X))=\frac{77}{27}$；(3)$MSE(\min(X))=\frac{224}{27}$
18. 样本方差的均方误差为 3.345。
19. 样本 $P(X<5)$ 的均方误差为 0.0133。
20. 样本的 0.3 分位点的均方误差为 1.34。

附录四 名词索引

参考文献

【1】Bowers，N. L. 著，余跃年，郑温瑜译：《精算数学》，上海科学技术出版社 1996 年版。

【2】Dick London 著，陈子毅译：《生存模型》，上海科学技术出版社 1995 年版。

【3】Dick London 著，徐诚浩译：《修匀数学》，上海科学技术出版社 1996 年版。

【4】Gerber H. G 著，成世学，严颖译：《数学风险论导引》，世界图书出版公司 1997 年版。

【5】Klugman S. A.，H. H. Panjer.，G. E. Willmot 著，吴岚译：《损失模型：从数据到决策》，人民邮电出版社 2009 年版。

【6】Rob Kaas，Marc Goovaerts，Jan Dhaene and Michel Denuit 著，唐启鹤、胡太忠、成世学译：《现代精算风险理论》，科学出版社 2005 年版。

【7】Ross，S. M. 著，王兆军、陈广雷、邹长亮译：《统计模拟》，人民邮电出版社 2007 年版。

【8】陈希儒：高等数理统计，中国科学技术大学出版社 1999 年版。

【9】黄向阳、金阳、肖宇谷等：《精算中常用的统计模型》，中国人民大学出版社 2009 年版。

【10】李晓林、孙佳美:《生命表基础》，中国财政经济出版社 2006 年版。

【11】李秀芳、傅安平:《寿险精算》，中国人民大学出版社 2002 年版。

【12】茆诗松、程依明、濮晓龙:《概率论与数理统计教程》，高等教育出版社 2004 年版。

【13】孟生旺、刘乐平：《非寿险精算学》，中国人民大学出版社 2007 年版。

【14】彭非、王伟：《生存分析》，中国人民大学出版社 2004 年版。

【15】严士健、刘秀芳:《概率与测度》，北京师范大学出版社 2003 年版。

【16】王晓军：《寿险精算学》，中国人民大学出版社 2005 年版。

【17】王星：《非参数统计》，清华大学出版社2009年版。

【18】吴岚、王燕：《风险理论》，中国财政经济出版社2006年版。

【19】肖争艳、高洪忠:《非寿险精算》，中国人民大学出版社2006年版。

【20】谢志刚、朱仁栋:《英汉精算学词汇》，上海科学技术出版社2000年版。

【21】张博：《精算学》，北京大学出版社2005年版。

【22】Bruno De Finetti: "*La Prision: ses lois logiques, ses sources subjectives*", 1937, Annales de l'Institut Henri Poincar.

【23】Anderson, T, W., and D. A. Darling, "*A Test of Goodness of Fit*", JASA. 1954, 49: 765 –769.

【24】Aarts and Korst, "*Simulated Annealing and Bolltzmann Machines*", Wiley, New York, 1989.

【25】Bowers, N., H., et al. "*Actuarial Mathematics*", Schaumburg IL: Society of Actuaries, 1986.

【26】Bowers, N. L. et al., "*Actuarial Mathematics*", Schaumburg, Illinois: Society of Actuaries. 1997.

【27】Bowman, A. W., A. Azzalini. "*Applied Smoothing Techniques for Data Analysis*: The Kernel Approach with S-Plus Illustrations", Oxford University Press, Oxford, 1997.

【28】Breslow, N. and J. Crowley, "*A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship*", Ann. Statist. 1974, 2: 437 –453.

【29】Broffit, J. D., "*Maximum Likelihood Alternatives to Actuarial Estimators of Mortality Rates*", TSA, 1984, XXXVI: 77 –142.

【30】Brown, R. L., "*Introduction to the Mathematics of Demography*," Winsted: ACTEX Publications, Inc., 1993.

【31】Dropkin, L. "*Some Considerations on Automobile Rating System Utilizing Individual Driving Records*", Proc. of the Casualty Actuarial Society, XLVI, 1959, 391 –405

【32】Fan, J., Q. Yao. "*Nonlinear Time Series: Parametric and Nonparametric Methods*", Springer, Princeton, 2005.

【33】Gerber, H. G., "*An Introduction to Mathematical Risk Theory*", University of Pennsylvania, Philadelphia: S. S. Huebner Foundation, 1980.

【34】Gompertz, B. "*On the nature of the function expressive of the law of human mortality*", Phil Trans Royal Soc, London 1825, 115: 513 –583.

【35】 Harold R. Greenlee, Jr. , and Alfonso D. Keh "*The* 1971 *Group Annuity Mortality Table*", Transactions Of Society Of Actuaries, 1971, 23: 569 – 622.

【36】 Hoem, J. M. , "*A flaw in Actuarial Exposed-to-risk Theory*", Scandinavian Actuarial Journal, 1984, 187 – 194.

【37】 Kaplan E. L and P. Meier, "*Nonparametric Estimation from Incomplete Observations*", JASA, 1958, 53: 457 – 481.

【38】 Kimeldorf G. S. and Jones D, "*A Bayesian Graduation*", TSA, 1967, XIX: 66 – 112.

【39】 Klugman, S. T. , H. H. Panjer and G. . E. Willmot "*Loss Models: From Data to Decisions*", John Wiley & Sons 1998.

【40】 London, D. , "*Survival Models and Their Estimation*", ACTEX Publications, Winsted, Connecticut, 1997.

【41】 Lehmann E. L. , "*Theory of Point Estimation*", John Wiley&Sons, 1983.

【42】 London, D. , "*Graduation: the Revision of Estimates*," Winsted: ACTEX Publications, Inc. , 1985.

【43】 Lowrie, W. B. , "*An Extension of Whittaker-Henderson Method of Graduation*", TSA, 1982, XXXIV: 329 – 72.

【44】 Makeham, W. M. , "*On the Law of Mortality and the Construction of Annuity Tables*" . J. Inst. Actuaries and Assur. Mag. 1860, 8: 301 – 310.

【45】 Nelson, W. A. , "*Theory and Applications of Hazard Plotting for Censored Failure Data.*" Technometrics 1972, 14: 945 – 966.

【46】 Panjer, H. H. and G. E. Willmot, "*Insurance Risk Models*", Schaumburg, Illinois: Society of Actuaries. 1992.

【47】 Press, S. J. , "*Bayesian Statistics: Principles, Models, and Applications*", John Wiley & Sons, 1989.

【48】 Ramsay, C. M. , "*Loading gross premiums for risk without using utility theory*", Transactions, 1987, Vol. XLV, 305 – 348.

【49】 Ramsey, F. P. , "*Truth and Probability*" written 1926, Published 1931 in Foundations of Mathematics and Other Logical Essays, Ch. VII, p. 156 – 198. Edited by R. B. Braithwaite. London: Kegan, Paul, Trench, Trubner & Co. Ltd. New York: Harcout, Brace and Company.

【50】 Schuette, D. R. , "*A Linear Programming Approach to Graduation*", TSA, 1978, XXX: 407 – 31.

【51】 Skellam, J. G. , "*A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets*

of Trials", J. R. Statistics. Soc. 1948, B10, 257 – 261.

【52】 Society of Actuaries Committee on Actuarial Principles, "*principles of actuarial science*", Transactions of the Society of Actuary, 1992, XLIV: 565 – 628.

【53】 Stephens, M. A., "*EDF Statistics for Goodness of Fit and Some Comparisons*", Journal of the American Statistical Association, 1974, 69: 730 – 737.

【54】 Stephens, M., "*Tests based on EDF statistics*", in D'agostino, R. and Stephens, M., eds, Goodness-of-Fit Techniques, New York: Marcel Dekker, 1986: 97 – 193.

【55】 Tenenbein, A. and Vanderhoof, I. T., "*New Mathematical Laws of Select and Ultimate. Mortality*," TSA, 1980, XXXII: 119 – 183.

【56】 Whittaker, E. T., "*On a New Method of Graduation*", Proc. Edin. Math. Soc., 1923, XLI: 63 – 75.

特别鸣谢

中国精算师资格考试用书《精算模型》得以顺利出版，得到了部分保险公司、高等院校及精算咨询机构的鼎立相助，在此特别鸣谢以下单位与个人（排名不分先后）：

单位：中央财经大学中国精算研究院

个人：李晓林　史　森　郭程宁　袁冬梅　钟　颖　杨东风
张　昊　王　浩　王义川　王立备

中国精算师协会

2010 年 11 月